Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 5

Using Importance Sampling to Improve Simulation in Linkage Analysis

Lars Ängquist* Ola Hössjer†

*University of Lund, Sweden, larsa@maths.lth.se

†University of Stockholm, Sweden, ola@math.su.se

Using Importance Sampling to Improve Simulation in Linkage Analysis*

Lars Ängquist and Ola Hössjer

Abstract

In this article we describe and discuss implementation of a weighted simulation procedure, importance sampling, in the context of nonparametric linkage analysis. The objective is to estimate genome-wide p-values, i.e. the probability that the maximal linkage score exceeds given thresholds under the null hypothesis of no linkage. In order to reduce variance of the estimate for large thresholds, we simulate linkage scores under a distribution different from the null with an artificial disease locus positioned somewhere along the genome. To compensate for the fact that we simulate under the wrong distribution, the simulated scores are reweighted using a certain likelihood ratio. If the sampling distribution are properly chosen the variance of the corresponding estimate is reduced. This results in accurate genome-wide p-value estimates for a wide range of large thresholds with a substantially smaller cost adjusted relative efficiency with respect to standard unweighted simulation.

We illustrate the performance of the method for several pedigree examples, discuss implementation including the amount of variance reduction and describe some possible generalizations.

KEYWORDS: Nonparametric linkage analysis, importance sampling, change of probability measure, exponential tilting, marker information, variance reduction, cost adjusted relative efficiency, genome-wide significance.

---

*The authors are grateful for comments made by two reviewers on an earlier version of this manuscript. Moreover, this research is sponsored by the Swedish Research Council, under contracts 6152-8013 and 621-2001-3288. Financial support is also given by the Wallenberg Laboratory, Department of Endocrinology, Malmö University Hospital, Lund University, Malmö.

1 Introduction

This article focuses on *simulation* and *nonparametric linkage (NPL) analysis*. The latter is a subfield of linkage analysis and it may be used to perform genome-wide scans designed to facilitate the search for genetic linkage of certain phenotypes (e.g. diseases) to different chromosomal regions. In this context we will use a simulation technique called *importance sampling* (cf. Hammersley and Handscomb, 1964) to calculate the statistical significance (p-values) of maximum NPL scores. This is the probability that the maximum NPL score exceeds a given threshold T under the null hypothesis of no linkage. When the NPL score is maximized over many chromosomes, the resulting p-values are orders of magnitude larger than the corresponding pointwise p-values. It is therefore important to correct for multiple testing. The most straightforward method is to use direct Monte Carlo simulation, which gives unbiased estimates of the p-value and is consistent in the limit of many simulations. However, this method is often very slow, especially for small p-values, large families and incomplete marker data. To remedy this, analytical formulas have been developed, based on approximating the NPL scores by a Gaussian process (Feingold et al., 1993; Lander and Kruglyak, 1995), or a skewness adjusted/transformed Gaussian process (Teng and Siegmund, 1998; Tang and Siegmund, 2001; Ängquist and Hössjer, 2003a). These formulas are fast to compute, but still approximations of the true p-value, the quality of which depend heavily on the information content of marker data. Importance sampling is an alternative which, like direct Monte Carlo simulation, gives consistent p-value estimates in the limit of many simulations. It typically is much faster than direct Monte Carlo simulation, especially for small p-values.

An example of importance sampling implemented for maximum NPL scores based on affected sib pairs is Malley et al. (2002) and further works using importance sampling in genetics and linkage analysis are e.g. Kong et al. (1992) and Cordell et al. (1995). For more information about linkage analysis in general cf. e.g. Ott (1999).

A brief guide to the general method now follows. The NPL score is simulated from a distribution $\tilde{P}$ which differs from the null hypothesis distribution P. Importance sampling is a weighted simulation technique, with weights depending on the likelihood ratio $L = d\tilde{P}/dP$. The choice of $\tilde{P}$ is crucial and should i) give an easily computed likelihood ratio ii) reduce the variance of the p-value estimate. In this paper we present a choice of $\tilde{P}$ based on simulating an artificial disease locus X along the predefined genomic region and then making inheritance vectors at X corresponding to high scores more likely according to an exponentially tilted distribution. We will consider linear combinations of p-value estimates based on differently tilted $\tilde{P}$ simulations. The distribution of the weights depends on the

threshold T. A good choice of $\tilde{P}$ will help us to estimate small p-values corresponding to large thresholds. This implies that a lower number of simulations will be needed to get a certain accuracy in the calculations, but the price to pay is that the simulation algorithm will be more time consuming per iteration. The efficiency of the importance sampling will depend on the relation between these properties.

In *Section 2* we introduce and present nonparametric linkage analysis and importance sampling in order to give the reader enough tools to follow the subsequent parts of the article. Next, in *Section 3* the suggested importance sampling method for perfect marker data is defined and p-value estimates from different $\tilde{P}$ measures are weighted to obtain maximal variance reduction. By perfect marker data we mean a dense set of markers, unrelated founders and that sufficiently many pedigree members are genotyped so that the score function S for each family can be unambigously determined at all loci. A generalization to incomplete marker information is given in *Section 4*. The results are presented in *Section 5*. A number of genome-wide scans, w.r.t. the 22 human autosomes based on four pedigree sets with one pedigree structure, are performed. We generally conclude that the procedure works well in the sense that it gives reliable p-values, even for very large NPL thresholds. In *Section 6* we briefly summarize the paper. Some technical details are given in the *appendices*.

2 Basic Theory

2.1 Definitions and Notation

Consider first one single pedigree including n related individuals; f founders and $(n-f)$ nonfounders. The allelic inheritance for a single pedigree, i.e. the distribution of the founder's alleles among the nonfounders, is based on the $m = 2(n-f)$ distinct meioses.

Following Donnelly (1983) the inheritance at locus x for a single pedigree is described by the *inheritance vector*, $v(x)$, defined as

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

In (1) p_i (m_i) equals 0 if the i^{th}nonfounder's paternal (maternal) allele originates from the grandfather and 1 if it originates from the grandmother.

Available computer programs with (nonparametric) linkage analysis implemented are e.g. GENEHUNTER (Kruglyak et al., 1996), ALLEGRO (Gudbjartsson et al., 2000) or MERLIN (Abecasis et al., 2002).

2.2 The NPL score

The *score function* is a function of the inheritance vector and measures compatibility between inheritance and phenotypes at a locus. Throughout this article we will use the S_{all} function (cf. Whittemore and Halpern, 1994). More information and discussions about the performance of score functions in general and S_{all} in particular may be found e.g. in Whittemore and Halpern (1994), Kruglyak et al. (1996), McPeek (1999) and Sengul et al. (2001).

As linkage measure we will use the *NPL score* (cf. Kruglyak et al., 1996; Kong and Cox, 1997) which for the k^{th}pedigree is defined as

$$Z_k(x) = \sum_w P(v_k(x) = w \mid \text{MD}_k) S_k(w), \tag{2}$$

where $P(v_k(x) = w \mid \text{MD}_k)$ is the probability function (at position x) for the inheritance vector $v_k(x)$ given the marker data MD_k and $S_k(\cdot)$ is the normalized one-locus score function,

$$S_k(v) \leftarrow \frac{S_k(v) - \mu_k}{\sigma_k}, \tag{3}$$

where $\mu_k = \sum_w S_k(w) p_{v_k}(w)$, $\sigma_k^2 = \sum_w S_k(w)^2 p_{v_k}(w) - \mu_k^2$ is the mean and variance of S_k before normalization, $p_{v_k}(w) = 2^{-m_k}$ is the probability distribution of $v_k(x)$ under the null hypothesis of no linkage and m_k is the number of meioses of pedigree k. Throughout this article we will assume that all score functions $S_k(\cdot)$ are normalized to have zero mean and unit variance, as in (3).

For a pedigree set with N distinct pedigrees and possibly non-equal pedigree weights, the NPL score is expressed as

$$Z(x) = \sum_{k=1}^{N} \gamma_k Z_k(x), \tag{4}$$

where $Z_k(x)$ is the NPL score in (2) assigned to the k^{th}pedigree, γ_k is the corresponding weight and $\sum_{k=1}^N \gamma_k^2 = 1$.

Equations (2)-(4) imply that $E\big(Z(x)\big) = 0$ and $V\big(Z(x)\big) \leq 1$ under H_0, with equality $V\big(Z(x)\big) = 1$ for perfect marker data. This implies that the perfect marker assumption leads to conservative tests, cf. Kruglyak et al. (1996). For large N and (close to) perfect marker information the NPL score may be approximated by a $N(0,1)$ normal distribution. The weights may be chosen according to different optimality criteria, cf. e.g. Sham et al. (1997), McPeek (1999) and Hössjer (2003a).

The maximum NPL score found during the analysis is formally expressed as

$$Z_{\max} = \sup\{Z(x),\ x \in \Omega\}, \tag{5}$$

where Ω is the chromosomal region(s) of interest in the study. This random variable $Z_{\max}$ is extensively used when we discuss issues of statistical significance, w.r.t. possible genetic linkage, below.

2.3 Significance

The p-value corresponding to the maximum NPL score $Z_{\max}$ is defined as

$$\alpha(z_{\max}) = P(Z_{\max} \geq z_{\max}), \tag{6}$$

where P denotes probability under the *null hypothesis* H_0 that no $x \in \Omega$ is linked to the disease locus. It tells us how probable it is, under H_0, to find a maximal NPL score greater than or equal to the observed $z_{\max}$.

When exact calculation of this p-value is not feasible one has to approximate it using simulation techniques or asymptotic approximation formulas. In this work we will use an importance sampling simulation method. Another approach, based on extreme value theory for Gaussian processes, is discussed by e.g. Lander and Botstein (1989), Feingold et al. (1993), Lander and Kruglyak (1995) and Tang and Siegmund (2001). Their findings were generalized to arbitrary pedigrees and adjusted for nonnormality in Ängquist and Hössjer (2003a).

2.4 Simulations

We approximate the p-value $\alpha(T)$ for a given threshold T by Monte Carlo simulation as follows. Generate J independent identically distributed (i.i.d.) replicates $Z^1_{\max}, \ldots, Z^J_{\max}$ of the random variable $Z_{\max}$ under H_0 and consider the estimate

$$\hat{\alpha}(T) = \frac{1}{J} \sum_{i=1}^{J} I(Z^i_{\max} \geq T), \tag{7}$$

where $I(A)$ is the indicator function for outcome A.

We assume absence of chiasma interference and that inheritance on different chromosomes is independent. To begin with, we assume perfect marker data. This assumption reduces the complexity of the $Z^i_{\max}$ computation since no hidden Markov algorithm is needed for evaluating the conditional inheritance distribution in (2). Then, in Section 4, we will relax this assumption.

For perfect marker data, each family score has the form $Z_k(x) = S_k\big(v_k(x)\big)$. To simulate $\{v_k(x); x \in \Omega\}$, each component of v_k may be seen as an independent and stationary Markov process with two states - 0 and 1 - and intensity matrix

$$\begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}. \tag{8}$$

This implies that jumps occur according to a Poisson process with intensity λ. The simulated inheritance vectors will give us the pedigree scores which facilitates, using (4), computation of the total NPL score. When map distance is measured in Morgans we have $\lambda = 1$.

For further information about simulation in the context of human linkage analysis cf. e.g. Boehnke (1986), Ploughman and Boehnke (1989), Ott (1989) and Terwilliger et al. (1993).

2.5 Importance Sampling

Let us give a brief introduction to importance sampling and how to use it for estimating $\alpha(T)$. Assume first that $\Omega = [0,l]$ consists of one chromosome, where l is the chromosome length. Let $Z = \{Z(x);\ x \in \Omega\}$ be the collection of all NPL scores on Ω and $f(Z) = I(Z_{\max} \geq T)$. Then

$$\alpha = \alpha(T) = E\big(f(Z)\big) = \int f(Z)dP(Z), \tag{9}$$

where E denotes expectation under the P distribution of no linkage. This value may be Monte Carlo estimated without bias, using (7), as $\hat{\alpha} = \frac{1}{J}\sum_{i=1}^{J} f(Z_i)$, where $\{Z_i\}$ are i.i.d. copies of Z. A possible improvement may be introduced by changing the probability measure from P to $\tilde{P}$ and considering the formula

$$\alpha = \alpha(T) = E\big(f(Z)\big) = \tilde{E}\big(L^{-1}(Z)f(Z)\big) \tag{10}$$

with $L(Z) = d\tilde{P}(Z)/dP(Z)$ the likelihood ratio. Formula (10) requires that $d\tilde{P}(z) > 0$ whenever $f(z)dP(z) > 0$, i.e. that the support of $\tilde{P}$ is a least as large as that of fP. By proper choice of $\tilde{P}$, we get a variance reduction when estimating α by

$$\tilde{\alpha} = \frac{1}{J}\sum_{i=1}^{J} L^{-1}(Z_i)f(Z_i), \tag{11}$$

where $\{Z_i\}$ are i.i.d. copies of Z under $\tilde{P}$.

An optimal choice of $\tilde{P}$ is

$$d\tilde{P}(Z) = \frac{f(Z)}{E[f(Z)]}dP(Z) = \frac{f(Z)}{\alpha}dP(Z), \tag{12}$$

which gives $L^{-1}(Z_i)f(Z_i) \equiv \alpha$ for all i and hence $V(\tilde{\alpha}) = 0$. This choice is not possible to use in practice though, since $\alpha = E(f(Z))$ is unknown. However, this may serve as a guidance to choose $\tilde{P}$ as close to (12) as possible. For instance, in the context of estimating p-values corresponding to exceedance of large thresholds, it might be useful to increase the probability for Z such that $f(Z)$ is large.

For further details on importance sampling cf. e.g. Hammersley and Handscomb (1964), Kotz and Johnson (1983) and Ross (2002).

3 Importance Sampling for Perfect Marker Data

3.1 One Single δ Value

Firstly, we describe the importance sampling algorithm in more detail for perfect marker data and one chromosome, $\Omega = [0, l]$. The estimator for several chromosomes is then defined in Section 3.3. The NPL score (4) for perfect marker data becomes

$$Z(x) = \sum_k \gamma_k S_k\big(v_k(x)\big), \tag{13}$$

where S_k is the score function and $v_k(x)$ the inheritance vector of pedigree k at locus x. We will sample from a probability measure

$$d\tilde{P}(z) = \frac{\int_\Omega g\big(z(x)\big)dx}{B} dP(z), \tag{14}$$

where g is a non-negative function, $B = \int_\Omega E\Big(g\big(Z(x)\big)\Big)dx$ a normalization constant and $z = \{z(x);\ x \in \Omega\}$ is the observed chromosome-wide NPL score. From this representation we get a likelihood ratio $L(z) = \int_\Omega g\big(z(x)\big)dx/B$ which is straightforward to compute as a function of z. We wish to choose g in such a way that $\tilde{P}$ in (14) reduces the variance of the estimator $\tilde{\alpha}$ as much as possible and, at the same time, construct a feasible sampling algorithm. Notice that (14) can be written as a mixture

$$d\tilde{P}(z) = \int_\Omega p(x) d\tilde{P}_x(z) dx, \tag{15}$$

where $d\tilde{P}_x(z) = g\big(z(x)\big)dP(z)/E\Big(g\big(Z(x)\big)\Big)$ and $p(x) = E\Big(g\big(Z(x)\big)\Big)/B$. This suggests a sampling algorithm:

Algorithm 1 *(Sampling Algorithm)*

1. *Generate X from the density* $p(\cdot)$.
2. *Conditionally on X, generate Z from* $\tilde{P}_X(\cdot)$. ■

Since Z is a stationary process under H_0, the expected value $E\Big(g\big(Z(x)\big)\Big)$ is independent of x and hence $p(x) = 1/l$ is the uniform distribution on Ω in *Step 1*. A possible choice of g is $g(y) = I(y \geq T)$, see e.g. Frigessi and Vercellis (1985) and Naiman and Priebe (2001). For imperfect marker data, it has recently been applied to p-value calculation for genome-wide linkage by Malley et al. (2002). An advantage of this g is that $\tilde{P}$ puts all it's probability mass on the set $\{Z_{\max} \geq T\}$ where f is positive. On the other hand, it is difficult to construct an exact sampling

algorithm for $\tilde{P}_X(\cdot)$ in *Step 2*. Malley et al. (2002) suggest an approximate fast algorithm based on the assumption that Z is a Gaussian process. This assumption is motivated by the central limit theorem for large N but can be quite inaccurate if N is small and/or there are a few large pedigrees in the data set.

In this paper, we don't approximate the NPL score by a Gaussian process. The main idea is to use an exponentially tilted density function for constructing $\tilde{P}$ (Ross, 2002, Section 8.5). This results in an exact and explicit simulation procedure for $\tilde{P}_x$, which gives unbiased estimates of α for arbitrary family structures, score functions S and weighting schemes γ. Let $\delta \geq 0$ be a given design parameter and put $g(y) = \exp(\delta y)$. Then $E\Big(g\big(Z(x)\big)\Big) = E\Big(\exp\big(\delta Z(x)\big)\Big) = M(\delta)$ is the moment generating function of $Z(x)$ under the null hypothesis of no linkage, evaluated at δ. One may note that $\delta = 0$ corresponds to standard simulation from P, i.e. under H_0. The likelihood ratio becomes

$$L(z) = \int_\Omega \exp\big(\delta z(x)\big) p(x) dx / M(\delta). \tag{16}$$

The crucial part of Step 2 in Algorithm 1 is to recognize that

$$\tilde{P}_X\big(Z(X) = z(X)\big) \propto \exp\big(\delta z(X)\big) \propto \prod_{k=1}^{N} \exp\Big(\delta \gamma_k S_k\big(v_k(X)\big)\Big), \tag{17}$$

since this makes it possible to simulate $Z(X)$ under $\tilde{P}_X$ by independently generating all inheritance vectors $v_k(X)$ according to the exponentially tilted distribution $\tilde{P}_X(v_k(X) = w) \propto \exp\big(\delta \gamma_k S_k(w)\big)$. Then, by the Markov property, $v_k(\cdot)$ is generated at all other loci. In more detail, Step 2 can be decomposed into three steps as follows:

Algorithm 2 *(Simulating NPL Score from $\tilde{P}_X$)*

> ***2a.*** *Independently for $k = 1, \ldots, N$, generate $v_k(X)$ from distribution $\tilde{P}_X(v_k(X) = w) = 2^{-m_k} \exp\big(\delta \gamma_k S_k(w)\big) / M_k(\delta)$ and $M_k(\delta) = 2^{-m_k} \sum_w \exp\big(\delta \gamma_k S_k(w)\big)$ is the moment generating function of $\gamma_k S_k\big(v_k(X)\big)$ under H_0.*
> ***2b.*** *Independently for $k = 1, \ldots, N$ do the following: Conditionally on $v_k(X)$ generate $v_k = \{v_k(x); x \in \Omega\}$ according to P.*
> ***2c.*** *Compute Z as in (13).* ■

In Appendix A, we show that these steps indeed give a valid sampling algorithm for $\tilde{P}_X(\cdot)$. For each k, *Step 2a* is just simulation from a discrete probability distribution and because of the assumption of no interference, *Step 2b* is obtained by generating crossovers to the left and right of X independently, see the discussion in Section 2.4.

For storage reasons it is more practical to loop over k, although we find the form above more useful when discussing the algorithm.

We may interpret X in *Step 1* as an artificial disease locus, uniformly positioned along Ω. Conditional on X, it is shown in Appendix C that $Z(X)$ approximately has a $N(\delta,1)$ distribution in the limit of large samples N. This gives a natural interpretation of δ as the asymptotic noncentrality parameter $\tilde{E}\big(Z(X)\big)$ (Feingold et al., 1993) at the artificial disease locus.

The exponentially tilted distribution of $Z(x)$ has previously been used (Kong and Cox, 1997) as an empirical likelihood, with δ as the unknown parameter at each locus x. Then, as a way to perform linkage analysis, δ is estimated as a function of x.

3.2 Weighting Estimates for Several δ Values

We consider estimates based on weighted averages. Introducing $0=\delta^1<\delta^2<\cdots<\delta^M$ and $w=(w_1,w_2,\ldots,w_M)$ we define

$$\tilde{\alpha}_w=\sum_{i=1}^{M} w_i\tilde{\alpha}_{\delta^i} \tag{18}$$

where $\tilde{\alpha}_\delta$ is the estimator (11) based on parameter δ and the weights satisfy $w_i\geq 0$ and $\sum_{i=1}^M w_i=1$. It follows that $E(\tilde{\alpha}_w)=\alpha$ and the variance

$$V(\tilde{\alpha}_w)=C_w(T)/J \tag{19}$$

is inversely proportional to the number of simulations J, with proportionality constant $C_w(T)=\sum_{i=1}^M w_i^2 C(T,\delta^i)$ and $C(T,\delta)=\tilde{E}\big((f(Z)/L(Z)-\alpha)^2\big)$. To minimize variance we use Tukey's inequality (Kotz and Johnson, 1988) to define the weights as

$$w_i\propto C(T,\delta^i)^{-1}. \tag{20}$$

In practice, we have to replace (20) by estimated weights, cf. Appendix D.

3.3 The Split-Merge Method

Although the importance sampling scheme works when Ω consists of several chromosomes we find it natural to utilize that marker data from different chromosomes is independent and split the estimation procedure into distinct estimates for the C chromosomes belonging to Ω and then merge this information into a joint genome-wide estimate

$$\tilde{\alpha}_w = \tilde{\alpha}_w(T) = 1 - \prod_{k=1}^{C} \left(1 - \tilde{\alpha}_{w,k}(T)\right), \tag{21}$$

where $\tilde{\alpha}_{w,k}(T)$ is the weighted estimate for the k^{th}chromosome.

4 Importance Sampling for Imperfect Marker Data

It is possible to generalize the importance sampling procedure to incomplete marker data. As in Section 3, we start with the case of one chromosome $\Omega = [0, l]$. Then, an estimate of α for several chromosomes is computed as in (21). By combining (2) and (4) we see that the NPL score $Z(\cdot)$ is a function of marker data $\text{MD} = (\text{MD}_1, ..., \text{MD}_N)$, which itself is a function of inheritance vectors $v = \{v_k, k = 1, ..., N\}$, where $v_k = \{v_k(x); x \in \Omega\}$, and founder genotypes at all marker loci for all pedigrees, MD_{found}. The simulation algorithm in Section 3 has to be modified, in that *Step 2c* is replaced by:

Algorithm 3 *(Simulating NPL Score Revisited)*

> ***2c1.*** *Generate founder genotypes $MD_{k,found}$, $\forall k$, at all marker loci.*
> ***2c2.*** *Generate MD_k, $\forall k$, as a function of v_k and $MD_{k,found}$ by segregating the founder alleles.*
> ***2c3.*** *Compute Z as function of MD by combining (2) and (4).* ■

For storage reasons, it is more practical to loop over k, although we prefer to write as above when discussing the algorithm. In order to get a manageable expression for the likelihood ratio, we define it as a function of MD rather than $Z(\cdot)$. It is shown in Appendix B that $L(\text{MD}) = \tilde{P}(\text{MD})/P(\text{MD})$ equals

$$L(\text{MD}) = \left(\int_\Omega p(x) \prod_{k=1}^{N} \sum_{w \in Z_2^{m_k}} \exp\left(\delta \gamma_k S_k(w)\right) P(v_k(x) = w | \text{MD}_k) dx \right) / M(\delta). \tag{22}$$

The importance sampling estimator of α is

$$\tilde{\alpha}(T) = \frac{1}{J} \sum_{i=1}^{J} L^{-1}(\text{MD}_i) f(\text{MD}_i), \tag{23}$$

where $\{\text{MD}_i\}_{i=1}^J$ are i.i.d. drawn from $\tilde{P}$ and $f(\text{MD}) = I(Z_{\max} \geq T)$. Estimates for several δ are combined in the same way as described in Section 3.2. The likelihood ratio (22) involves the conditional inheritance distributions $P(v_k(x) \mid \text{MD}_k)$

for each pedigree. This distribution is computationally involved for large pedigrees, cf. Kruglyak et al. (1996). Notice however that $P(v_k(x) \mid \mathrm{MD}_k)$ appears in (2) and hence has to be computed for all pedigrees at all loci in order to define $Z_{\max}$. Therefore, the *additional* computational burden to evaluate the likelihood ratio is relatively small.

5 Results

For simplicity of interpreting the effect of varying the pedigree structure, we considered only data sets with N pedigrees of the same kind, i.e. homogeneous pedigree sets, chosen from Figure 1. We expect results for mixed pedigree sets to have sim-

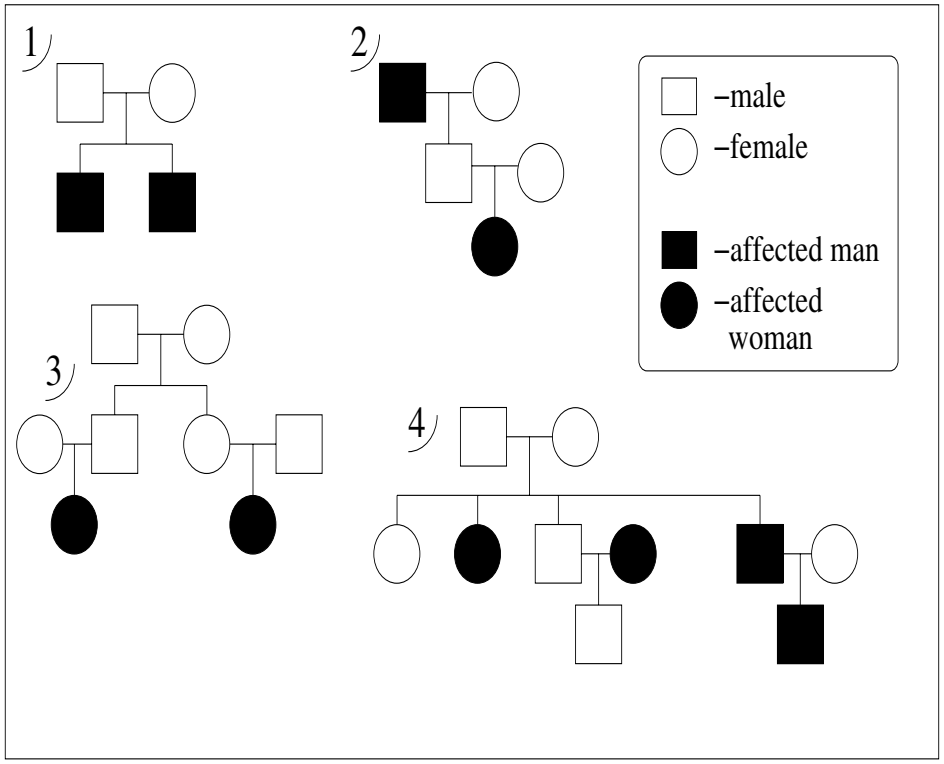

Figure 1: Four pedigrees of different structure.

ilar performance. Moreover, we use $N = 60$ pedigrees, equal pedigree weighting, i.e. $\gamma_k = 1/\sqrt{N}$ for $k = 1,\ldots,N$ and the grid $\{\delta^i\}_{i=1}^M = \{(i-1)\cdot 5.5/(M-1)\}_{i=1}^M$. We compare, for perfect marker data, the weighted simulations to traditional unweighted simulations and an approximation formula based on extreme value theory for stochastic processes (Ängquist and Hössjer, 2003a).

5.1 Cost Adjusted Efficiency

The importance sampling algorithm involves the design vector $\delta = (\delta^1, \delta^2, \ldots, \delta^M)$ and the corresponding weights $w = (w_1, w_2, \ldots, w_M)$, that should be chosen in order to reduce variance as much as possible. Recall that the importance sampling estimator (18) is unbiased with variance (19).

The advantages of this procedure should be balanced against the additional computational cost compared to traditional Monte Carlo simulation (i.e. $\delta = 0$). We define CR_i as the *cost ratio* for δ^i, i.e. the ratio of the computation time per simulation for importance sampling with $\delta = \delta^i$ and Monte Carlo simulation ($\delta = 0$). It depends to some extent on the implementation of the algorithm. We assume that CR_i is independent of T and define the *cost adjusted relative efficiency* (see also Malley et al., 2002) as

$$\mathrm{RE}(T) = C_{\mathrm{MC}}(T)/(C_w(T) * \sum_{i=1}^{M} \mathrm{CR}_i), \tag{24}$$

where $C_{\mathrm{MC}}(T) = C(T,0) = \alpha(T)\big(1-\alpha(T)\big)$, and $C_w(T)$ is defined in Section 3.2. This is the ratio of the computation time needed for the Monte Carlo estimate to attain the same variance as the importance sampling estimate. In our case $\mathrm{CR}_1 = 1$ since $\delta_1 = 0$ and $\mathrm{CR}_i = \mathrm{CR}$ is independent of i for $2 \le i \le M$. Therefore (24) reduces to

$$\mathrm{RE}(T) = C_{\mathrm{MC}}(T)/(C_w(T) * \big(1 + (M-1)\mathrm{CR}\big). \tag{25}$$

Because of the asymptotic normality of Z under H_0, we expect $\mathrm{RE}(T)$ to be fairly robust against variations of sample size, pedigree structures, weighting schemes and score functions. It is a bit more sensitive to variations in total genomic length and degree of marker informativity. For suboptimal choices of δ the estimate $\tilde{\alpha}$ has high variability. This is particularly true for $\delta = 0$ and large thresholds T. For this reason, we estimate $C_{\mathrm{MC}}(T)$ by $\hat{C}_{\mathrm{MC}}(T) = \tilde{\alpha}_w(1-\tilde{\alpha}_w)$, where $\tilde{\alpha}_w$ is the weighted importance sampling estimate. The estimate $\hat{C}_w(T)$ is defined in Appendix D.

Estimates of $\mathrm{RE}(T)$ for one chromosome and various choices of l and M are shown in Figure 2. The efficiency increases with T, e.g. for T=8 we get approximately $\mathrm{RE}(T) = 10^7$. The ratio decreases slowly with the genetic length l, despite that the cost ratio decreases as well ($\mathrm{CR} \approx 3$ when $l = 1$ and $\mathrm{CR} \approx 2$ when $l = 10$). As long as the δ-grid covers the set of thresholds $\mathbb{T}$ in the sense that some $C(T, \delta^i)$ is small for all $T \in \mathbb{T}$, there is no need to increase M further.

5.2 Displaying the Weights

Figure 3 displays the weight vector w as function of the threshold T for one chromosome of length 3.0 Morgans.

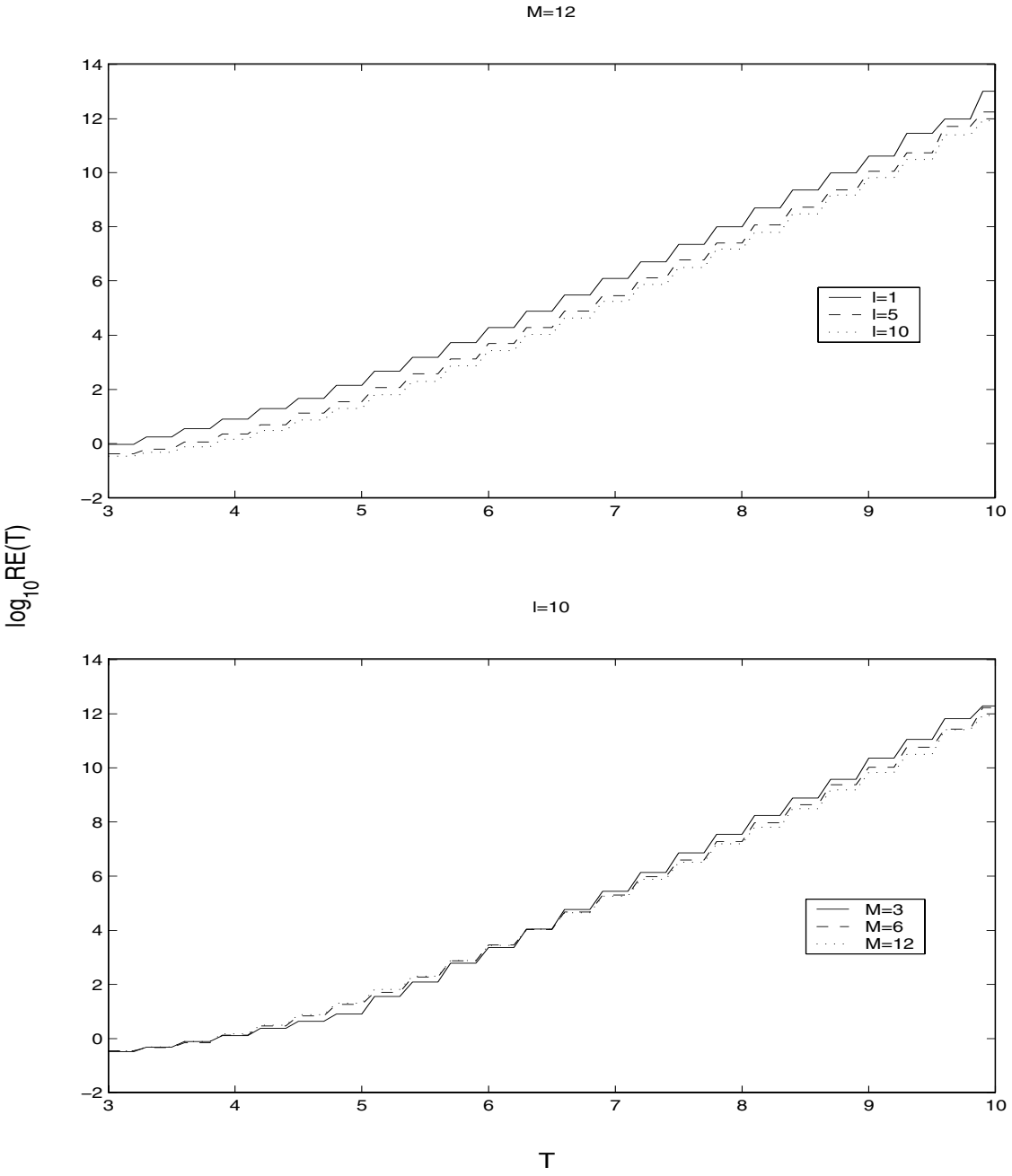

Figure 2: Relative efficiency estimates based on J=2500 simulations for Pedigree 3, with number of pedigrees $N = 60$, thresholds T, tuning constants, introduced in Appendix D, $\varepsilon = (0.001, 0.95)$, genome length l=1,5 or 10 Morgans and grid of tilting parameters $\{\delta^i\}_{i=1}^{M} = \{(i-1)\cdot 2.75\}_{i=1}^{3}$, $\{(i-1)\cdot 1.1\}_{i=1}^{6}$ or $\{(i-1)\cdot 0.5\}_{i=1}^{12}$ for M=3, 6 or 12 respectively, where M is the number of gridpoints. Due to implementation reasons we used $\mathrm{CR}_1 = \mathrm{CR} > 1$ which (with a reduced effect for increasing M) slightly underestimates $\mathrm{RE}(T)$ in (25).

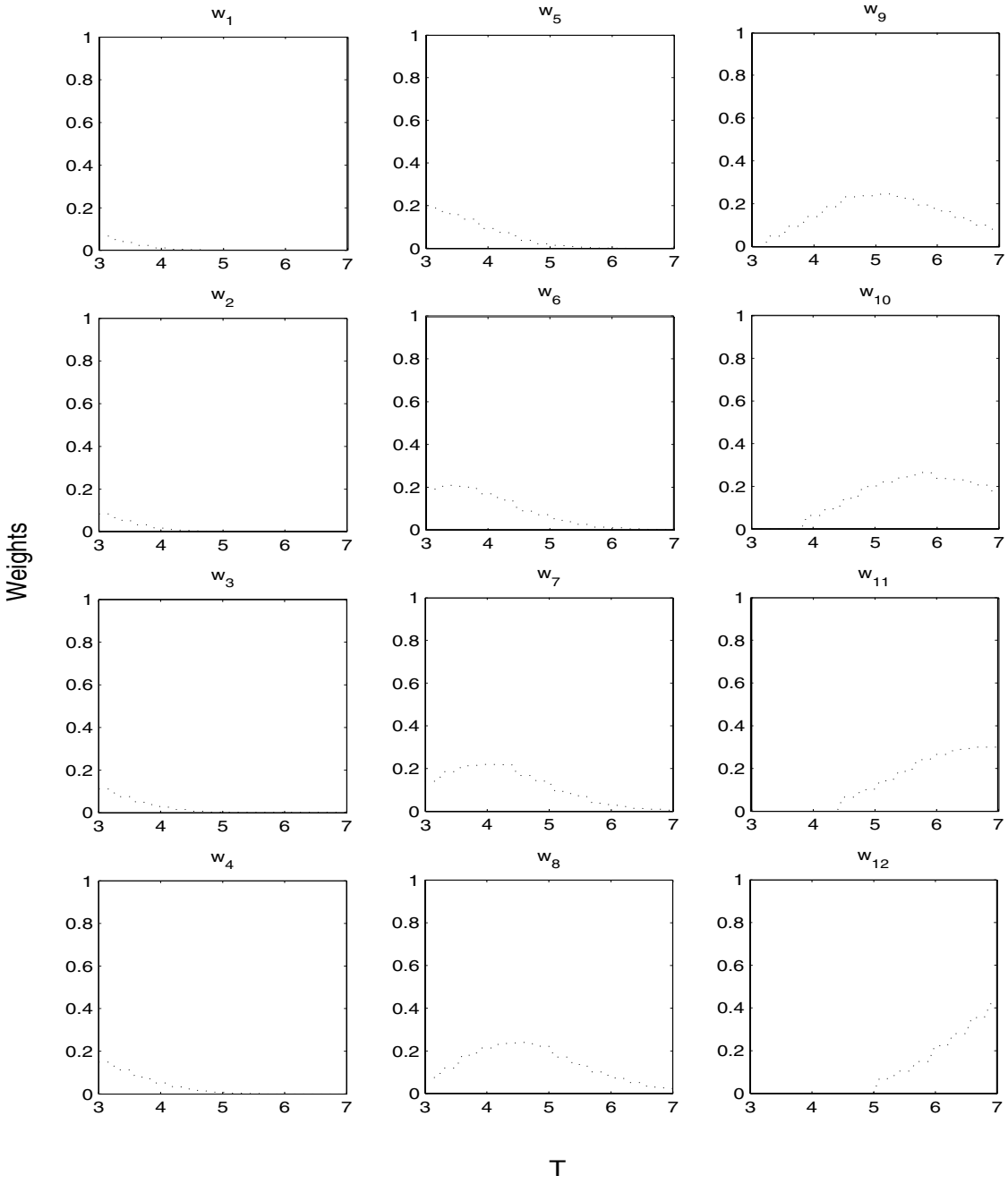

Figure 3: Illustration of the weight vector $w = (w_1, \ldots, w_{12})$ as function of the threshold T for Pedigree 3 when the number of simulations and pedigrees are J=10000 and N=60 respectively, the grid of tilting parameters $\{\delta^i\}_{i=1}^{12} = \{(i-1)\cdot 0.5\}_{i=1}^{12}$ and the tuning constants given in Appendix D are set to ε=(0.001, 0.95).

The procedure (41) gives large weights w_i to small i when T is small and to large i when T is large. This means that we successively include and remove new $\tilde{\alpha}_{\delta^i}$ when increasing T. With high probability new estimates $\tilde{\alpha}_{\delta^i}$ are included and removed in increasing δ^i order. One may notice that the number of simultaneously positive weights, for this setting, is about 6-10 (less for very small/large Ts). This corresponds to a range of δ roughly between 3 and 4.

5.3 Four Examples of Full-Scale Autosomal Investigations

To further test the simulation procedure, four pedigree sets (cf. Figure 1) are used in a full autosomal setting, i.e. the genome region consists of all the 22 human autosomes. We perform a split-merge analysis (Section 3.3) and the procedure is successful i.e. we are able to get good estimates for p-values corresponding to a wide range of thresholds (cf. Figure 4).

One may note that when using this technique it is possible to find estimates for much smaller p-values (larger thresholds) than when using traditional unweighted simulations. Importance sampling with $J = 3000$ and $M = 12$ leads to accurate estimates of p-values of magnitude less than 10^{-10}, whereas Monte Carlo simulation with $J = 100000$ leads to accurate p-value estimates down to 10^{-5}.

6 Discussion

In this article we have discussed a method to calculate genome-wide significance levels for arbitrary pedigree sets using importance sampling. The main strength of the method described, compared to traditional simulation techniques, is that it makes it possible to, given a reasonable number of simulations, accurately estimate quantitatively very small p-values. Alternatively, less simulations are needed to attain a given accuracy (i.e. variance) of the estimator. This may be important e.g. when searching for an overall measure of significance w.r.t. a lot of different genome scans with a large total genetic length.

We have generally assumed perfect data but have also (in Section 4) generalized the method to incomplete marker information. One reason for assuming perfect marker data in the simulations is the possibility to compare results to the approximation formula of Ängquist and Hössjer (2003a). The simulations for $N = 60$ pedigrees show good agreement between the two methods. (It can be shown that the same is true also for other values of N, since the analytical method adjusts for non-normality caused by lack of validity of the central limit theorem.) The approximation formula is faster to compute, whereas our importance sampling scheme gives exact p-values in the limit of many simulations. Moreover, it naturally extends

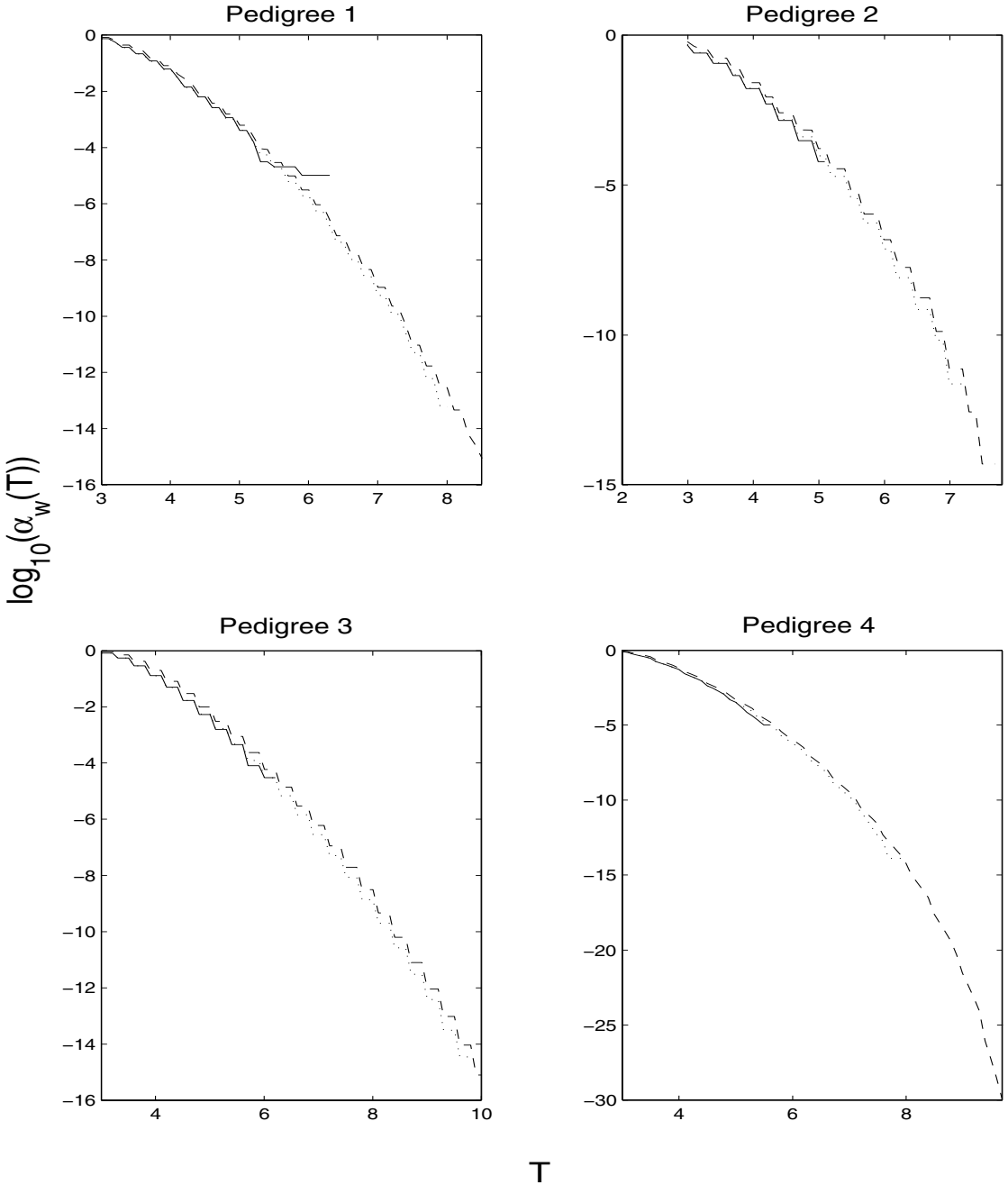

Figure 4: Significance comparisons between importance sampling with J=3000 simulations $(\cdots)$, unweighted simulation with J=100000 simulations (–) and an approximation formula $(--)$. The simulation is performed over all the 22 autosomes with a total chromosomal genetic length of 35.75 Morgans (Collins et al., 1996). Additional parameters: number of pedigrees N=60, thresholds T, the grid of tilting parameters $\{\delta^i\}_{i=1}^{12} = \{(i-1)\cdot 0.5\}_{i=1}^{12}$ and the tuning constants, defined in Appendix D, ε=(0.001, 0.95).

to incomplete marker data of any kind, whereas the analytical approximation formula is best suited for fully polymorphic markers with uniform marker spacing.

The importance sampling approach of Malley et al. (2002) is faster than ours, since it is based on Gaussian process approximations, thereby avoiding the need to simulate inheritance vectors. An advantage of our approach is that it gives unbiased estimates of the p-value, since it does not rely on approximating the NPL score by a Gaussian or transformed Gaussian process. It works for arbitrary combinations of pedigree structures (even inbred ones with loops), weighting schemes and score functions. This is true both for perfect as well as imperfect marker data, in the latter case regardless of marker spacing and heterozygosity of markers. The method of Malley et al. (2002) allows for finite marker spacing but requires each marker to be fully polymorphic. Further, their algorithm depends on the fact that the covariance function of Z is doubly exponential. This is certainly true when all families are affected sib pairs, but not for general collections of pedigrees (Hössjer, 2003b).

Note that in some cases, e.g. when calculating a single p-value, it may be preferable to choose $M = 1$ and use the unweighted estimate $\tilde{\alpha}_\delta$. Although simpler, this forces us to choose an appropriate δ. When considering a small number of thresholds and a δ-grid, a linear combination of the estimates from the closest surrounding δs may be used. For more details on this method, see Ängquist and Hössjer (2003b).

7 Colophon

7.1 Programs

All calculations have been performed using MATLAB. For editing and typesetting purposes we have used LaTeX and BibTeX.

7.2 Information About the Authors

- Lars Ängquist (Corresponding author, Ph.D.-student), Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University. Regular mail: Box 118, S-22100, Lund, Sweden. E-mail: larsa@maths.lth.se.

- Ola Hössjer (Professor), Department of Mathematics, Stockholm University, Stockholm, Sweden. E-mail: ola@math.su.se.

A Verifying Sampling Algorithm for Perfect Data

Given X, let $\tilde{P}_X$ be the distribution of Z that results when applying Algorithm 2. We will show that $\tilde{P}_X$ coincides with the definition earlier in Section 3, since this will imply that *Step 2* in Algorithm 1 can be replaced by Algorithm 2. We do this by showing that $L_X(z) = d\tilde{P}_X(z)/dP(z) = \exp\big(\delta z(X)\big)/M(\delta)$.

Let $v(x) = \{v_1(x), \ldots, v_N(x)\}$ and $v = \{v(x); x \in \Omega\}$ be the collection of inheritance vectors at locus x and at all loci respectively. Then

$$\frac{d\tilde{P}_X(v)}{dP(v)} = \frac{d\tilde{P}_X(v(X))}{dP(v(X))} = \prod_{i=1}^{N} \frac{\exp\Big(\delta\gamma_k S_k\big(v_k(X)\big)\Big)}{M_k(\delta)} = L_X(z), \tag{26}$$

where $z = \{z(x); x \in \Omega\}$ is defined as in (13). In the first equality we used *Step 2b*, from which it follows that $d\tilde{P}_X(v|v(X))$ and $dP(v|v(X))$ have the same distribution, and in the last equality we used $M(\delta) = \prod_{k=1}^{N} M_k(\delta)$.

It remains to verify that the same likelihood ratio is obtained when replacing v by z. First write $z = F(v)$ to indicate that z is a function of v and then define the set $A = A(z) = \{v; F(v) = z\}$. Then, it follows from (26) that

$$d\tilde{P}_X(z) = \int_A d\tilde{P}_X(v) = \int_A L_X(z)dP(v) = L_X(z)\int_A dP(v) = L_X(z)dP(z), \tag{27}$$

as was to be proved. The crucial step in the last equation is the constancy of $d\tilde{P}_X(v)/dP(v)$ over A, since this implies that $L_X(z)$ can be factored out from the integral.

B Verifying Sampling Algorithm for Imperfect Data

As noted in *Step 2c1*, the marker data is a function of v and the founder alleles at all marker loci for all pedigrees. We write MD $\sim v$ if marker data is consistent with v. That is, at all marker loci segregation of marker alleles is consistent with v at the corresponding loci and for all pedigrees. The requirement MD $\sim v$ is less stringent if markers have a low heterozygosity since then more segregation patterns are possible. Define $B = B(\text{MD})$ by $B = \{v\ ;\ \text{MD} \sim v\}$. Then

$$\tilde{P}(\text{MD}) = \int_B P(\text{MD} \mid v)d\tilde{P}(v). \tag{28}$$

By Bayes' rule, the conditional inheritance distribution of v given markers can be written as

$$dP(v \mid \text{MD}) = \frac{P(\text{MD} \mid v)dP(v)}{P(\text{MD})}. \tag{29}$$

From (16) and (26)-(27) it follows that

$$d\tilde{P}(v) = dP(v)\left(\int_{\Omega} p(x)\prod_{k=1}^{N}\exp\left(\delta\gamma_k S_k\big(v_k(x)\big)\right)dx\right)/M(\delta). \tag{30}$$

Inserting these two equations into (28) and changing the order of integration we get

$$\tilde{P}(\mathrm{MD}) = P(\mathrm{MD})\left(\int_{\Omega} p(x)\int_{B}\prod_{k=1}^{N}\exp\left(\delta\gamma_k S_k\big(v_k(x)\big)\right)dP(v \mid \mathrm{MD})dx\right)/M(\delta). \tag{31}$$

Because of independence of marker data for different pedigrees

$$dP(v \mid \mathrm{MD}) = \prod_{k=1}^{N} dP(v_k \mid \mathrm{MD}_k). \tag{32}$$

Therefore, the inner integral in (31) may be written as

$$\prod_{k=1}^{N}\sum_{w}\exp\big(\delta\gamma_k S_k(w)\big)P(v_k(x) = w \mid \mathrm{MD}_k). \tag{33}$$

Combining (31), (33) and taking the likelihood ratio we obtain (22).

C Asymptotic Distribution of $Z(X)$

We begin with deriving an approximation of the expected NPL score at locus X under the probability measure $\tilde{P}$ and perfect marker data. First consider the k^{th}pedigree only and introduce $\delta_k = \delta\gamma_k$. Then

$$\tilde{E}\big(Z_k(X)\big) = \sum_z z\tilde{P}(Z_k(X) = z) = \frac{\sum_z z\,P(Z_k(X) = z)\exp(\delta_k z)}{\sum_z P(Z_k(X) = z)\exp(\delta_k z)} = \frac{M_k'(\delta)}{M_k(\delta)}, \tag{34}$$

where $M_k(\delta) = 2^{-m_k}\sum_w \exp\big(\delta\gamma_k S_k(w)\big)$ (cf. Algorithm 2) and the derivative is taken with respect to δ_k. Now, using a Taylor expansion,

$$M_k(\delta) = 1 + \delta_k E\big(Z_k(X)\big) + \frac{\delta_k^2}{2}E\big(Z_k(X)^2\big) + o(\delta_k^2) = 1 + \frac{\delta_k^2}{2} + o(\delta_k^2) \tag{35}$$

and (34)-(35) imply $\tilde{E}\big(Z_k(X)\big) = M_k'(\delta)/M_k(\delta) = \delta_k + o(\delta_k)$ as $\delta_k \to 0$.

For the total linkage score we use that $\sum_{k=1}^{N} \gamma_k^2 = 1$ and find

$$\tilde{E}\big(Z(X)\big) = \sum_{k=1}^{N} \gamma_k \tilde{E}\big(Z_k(X)\big) = \sum_{k=1}^{N} \gamma_k \big(\delta_k + o(\delta_k)\big) = \delta + o(1). \tag{36}$$

as $N \to \infty$ and $\max_{1 \le k \le N} \gamma_k \to 0$.

Further, a similar calculation shows that $\tilde{V}\big(Z(X)\big) = 1 + o(1)$ under the same conditions. Then a Central Limit Theorem argument implies, under mild regularity conditions on the set of pedigree structures, that asymptotically

$$Z(X) \xrightarrow{\mathscr{D}} N(\delta, 1), \tag{37}$$

where $\xrightarrow{\mathscr{D}}$ denotes convergence in distribution under $\tilde{P}$.

D Computing the Weighted Estimate

To find the weights in (20) estimate $C(T, \delta)$ through

$$\hat{C}(T, \delta) = \sum_{i=1}^{J} (f(Z_i)/L(Z_i) - \tilde{\alpha}_\delta)^2 / J. \tag{38}$$

Notice that these weights depend on T whereas all $\tilde{P}$ do not. Hence it is just the weighting in (18) and the function f in (10) that depend on the threshold. Introduce

$$\beta_\delta(T) = \tilde{P}(Z_{\max} \ge T) \tag{39}$$

for the probability that the maximum NPL score exceeds T under $\tilde{P}$. We interpret $\beta_\delta(T)$ as the power of the test $Z_{\max} \ge T$ to detect the artificial disease locus at X under $\tilde{P}$. Moreover, define an estimator

$$\hat{\beta}_\delta(T) = \frac{1}{J} \sum_{i=1}^{J} I(Z_{\max,i} \ge T). \tag{40}$$

based on J i.i.d. maximal NPL scores $\{Z_{\max,i}\}$ under $\tilde{P}$.

One problem with the variance estimator is a tendency to be noisy for extreme values of (39). To avoid this effect we define a truncation rule of the weights, given T,

$$w_i \propto \hat{C}(T, \delta^i)^{-1} \qquad \text{if} \qquad \varepsilon_1 \le \hat{\beta}_{\delta^i}(T) \le \varepsilon_2, \tag{41}$$

with the complementary rule of putting the weights to 0 in all other cases, except for the situation where no $\hat{\beta}_\delta$-value satisfies the above inequalities. Then we perform traditional (non-weighted) simulation and put $w_1 = 1$.

The choice of $\varepsilon = (\varepsilon_1, \varepsilon_2)$ is not that crucial, see Ängquist and Hössjer (2003b) for more details.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, *30*, 97–101.

Ängquist, L. and Hössjer, O. (2003a). *Improving the calculation of statistical significance in genome-wide scans* (Tech. Rep. No. 2003:3). Lund: Department of Mathematical Statistics, Lund University. Under revision for *Biostatistics*.

Ängquist, L. and Hössjer, O. (2003b). *Using importance sampling to improve simulation in linkage analysis* (Tech. Rep. No. 2003:34). Lund: Department of Mathematical Statistics, Lund 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.

Collins, A., Frezal, J., Teague, J. and Morton, N. E. (1996). A metric map of humans: 23.500 loci in 850 bands. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 14771–14775.

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.

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

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.

Frigessi, A. and Vercellis, C. (1985). An analysis of Monte Carlo algorithms for counting problems. *Calcolo*, *22*(4), 413–428.

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.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo methods*. New York: John Wiley & Sons.

Hössjer, O. (2003a). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Annals of Statistics*, *31*(4), 1075–1109.

Hössjer, O. (2003b). *Spectral decomposition of score functions in linkage analysis* (Tech. Rep. No. 2003:21). Stockholm: Department of Mathematical Statistics, Stockholm University.

Kong, A. and Cox, N. (1997). Allele-sharing models: Lod scores and accurate

linkage tests. *American Journal of Human Genetics*, *61*, 1179–1188.

Kong, A., Frigge, M., Irwin, M. and Cox, N. (1992). Importance sampling. I. Computing multimodal *p*-values in linkage analysis. *American Journal of Human Genetics*, *51*, 1413–1429.

Kotz, S. and Johnson, N. L. (1983). Importance sampling. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia in statistical sciences* (Vol. 4, p. 25). A Wiley-Interscience Publication: John Wiley & Sons.

Kotz, S. and Johnson, N. L. (1988). Tukey's inequality for optimal weights. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia in statistical sciences* (Vol. 9, pp. 361–362). A Wiley-Interscience Publication: John Wiley & Sons.

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*, *55*, 1347–1363.

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

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

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.

Naiman, D. Q. and Priebe, C. (2001). Computing scan statistic *p*-values using importance sampling, with applications to genetics and medical image analysis. *Journal of Computational and Graphical Statistics*, *10*(2), 296–328.

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.

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

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.

Ross, S. M. (2002). *Simulation* (Third ed.). San Diego: Academic 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.

Sham, P., Zhao, J. and Curtis, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Annals of Human Genetics*, *61*, 61–69.

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

Teng, J. and Siegmund, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics*, *54*, 1247–1265.

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.