Mathematical Statistics
Stockholm University

Using Importance Sampling to Improve Simulation in Linkage Analysis

Lars Ängquist and Ola Hössjer

Research Report 2003:19

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 **2003:19**,
http://www.math.su.se/matstat

Using Importance Sampling to Improve Simulation in Linkage Analysis

Lars Ängquist*and Ola Hössjer†

December 2003

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 a given threshold under the null hypothesis of no linkage. In order to reduce variance of the p-value 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 design parameters of the sampling distribution are chosen correctly, the variance of the final significance value estimate is reduced. This results in more accurate genome-wide p-value estimates for large thresholds, based on a substantially smaller number of simulations than is needed using traditional unweighted simulation.

We illustrate the performance of the method for several pedigree examples, discuss implementation including choice of sampling parameters and describe some possible generalizations.

*Department of Mathematical Statistics, Lund University, Lund and Wallenberg Laboratory, Department of Endocrinology, Malmö University Hospital, Lund University, Malmö.

†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.

Key words: Nonparametric linkage analysis, importance sampling, change of probability measure, exponential tilting, marker information, genome-wide significance.

Contents

1 Introduction

This article focuses on *simulation* and *nonparametric linkage analysis* (NPL analysis; cf. e.g. Kruglyak et al. (1996). NPL analysis 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 (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. Other examples of importance sampling being used in the context of genetics and linkage analysis can be found e.g. in Kong et al. (1992) and Cordell et al. (1995).

A short description of the main theme of the importance sampling technique is as follows. The NPL score is simulated from a distribution P_δ which differs from the null hypothesis distribution P_0. Importance sampling is a weighted simulation technique, with weights depending on the likelihood ratio $L = dP_0/dP_\delta$. The choice of P_δ 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 P_δ based on simulating an artificial disease locus x_0 along the predefined genome Ω and then making inheritance vectors at x_0 corresponding to high scores more likely according to an exponentially tilted distribution. The choice of the tilting parameter δ is discussed in detail. We will consider linear combinations of p-value estimates based on differently tilted P_δ simulations. For different thresholds T, different δ values are assigned high weights. A good choice of weights 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 advanced and that each simulation will be more time consuming. The efficiency of the importance sampling will depend on the relation between these properties. For more information about linkage analysis in general cf. e.g. Ott (1999) and an example of a different permutation test method for computing p-values is described by Abney at al. (2002) in the QTL case.

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* we first define the basic

version of the suggested importance sampling method for perfect marker data and then discuss several possible alternatives, refinements and improvements of this method. A generalization to handle incomplete marker information is given in *section 4*. The results are presented in *section 5*. We perform a number of specific investigations, in order to understand the performance of the importance sampler, in particular how to weight p-value estimates from different P_δ measures. We then estimate p-values for genome-wide scans based on four different pedigree sets w.r.t. all 22 human autosomes. 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 present nonfounders, is based on the $m = 2(n-f)$ distinct meioses.

Following Kruglyak et al. (1996) 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 which facilitates the performance of (non-parametric) linkage analysis 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 (Whittemore and Halpern, 1994) which is defined as

$$S_{all}(v) = 2^{-A} \sum_{h} \prod_{i=1}^{2f} b_i(h)! \tag{2}$$

where A is the number of affected individuals in the pedigree, h is a selection that picks one out of the two present alleles for each affected individual, $b_i(h)$ is the number of copies of founder allele i ($i = 1, 2, \ldots, 2f$) in h (given v) and the sum includes terms corresponding to all the 2^A distinct ways of choosing h. 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 a linkage measure we will use the *NPL score* (Kruglyak et al., 1996) which, for a single family, is defined as

$$\bar{Z}(x) = \sum_w P(v(x) = w \mid \text{MD}) Z(w), \tag{3}$$

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

$$Z(v) = \frac{S(v) - \mu}{\sigma}, \tag{4}$$

where $\mu = \sum_w S(w) p_v(w)$, $\sigma^2 = \sum_w S(w)^2 p_v(w) - \mu^2$ and $p_v(w) = 2^{-m}$ is the probability distribution of $v(x)$ under the null hypothesis (H_0) of no linkage.

The purpose of the markers is to give as much information as possible regarding the inheritance vector at a given locus. The information measure of Kruglyak et al. (1996) is a function of the entropy of the distribution $P(v(x) \mid \text{MD})$. It ranges between 0 (no marker information) and 1 (perfect marker information). In this paper we mainly consider the case of fully informative markers with $\bar{Z}(x) = Z(v(x))$ in (3). However, in section 4 we also discuss the case of incomplete marker information.

Generalized to a set of N distinct pedigrees, with the additional possibility of non-equal pedigree weighting, our final NPL score (cf. Kruglyak et al., 1996; Kong and Cox, 1997) is expressed as

$$\bar{Z}(x) = \frac{\sum_{k=1}^N \gamma_k \bar{Z}_k(x)}{\sqrt{\sum_{k=1}^N \gamma_k^2}}, \tag{5}$$

where $\bar{Z}_k(x)$ is the NPL score in (3) assigned to the k^{th}pedigree and γ_k is the corresponding weight.

Equations (3)-(5) imply that $E\Big(\bar{Z}(x)\Big) = 0$ and $Var\Big(\bar{Z}(x)\Big) \leq 1$ under H_0, with equality $Var\Big(\bar{Z}(x)\Big) = 1$ 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 (2003). If weights depend on family scores at a different locus (loci), located on other chromosomes, we call the approach a *conditional* two-locus (multi-locus) NPL-analysis (cf. e.g. Cox et al., 1999 and Ängquist, 2001).

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

$$\bar{Z}_{\max} = \sup\{\bar{Z}(x),\ x \in \Omega\}, \tag{6}$$

where Ω is the chromosomal region(s) of interest in the study. This random variable $\bar{Z}_{\max}$ 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 $\bar{Z}_{\max}$ is defined as

$$\alpha(\bar{z}_{\max}) = P(\bar{Z}_{\max} \geq \bar{z}_{\max} \mid H_0), \tag{7}$$

which tells us how probable it is to find a maximal NPL score greater than or equal to the observed $\bar{z}_{\max}$ under the *null hypothesis* (H_0) that no $x \in \Omega$ is linked to the disease locus.

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

2.4 Simulations

Here we describe the basic method of Monte Carlo simulation applied to the situation of generating NPL scores.

To approximate the p-value $\alpha(T)$ for a given threshold T we generate J independent identically distributed (i.i.d.) replicates $\bar{Z}^1_{\max}, \ldots, \bar{Z}^J_{\max}$ of the random variable $\bar{Z}_{\max}$ under the null hypothesis of no linkage and then we consider the estimate

$$\hat{\alpha}(T) = \frac{1}{J} \sum_{i=1}^{J} I(\bar{Z}^i_{\max} \geq T), \tag{8}$$

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

We simulate under H_0, assume no chiasma interference is present and that inheritance on different chromosomes is independent. To begin with, we assume perfect marker data. This assumption reduces the complexity of the $\bar{Z}^i_{\max}$ computation since no hidden Markov algorithm is needed for evaluating the conditional inheritance distribution in (3). Then, in section 4, we will relax this assumption.

Each single component of the inheritance vector of a pedigree may be seen as an independent and stationary Markov process with two states - 0 and 1 - and intensity matrix

$$\begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}. \tag{9}$$

This implies that jumps occur according to a Poisson process with intensity λ. The simulated inheritance vectors will give us the family scores which facilitates, using (5), the computation of the total NPL score.

Measuring the map distance in Morgans we get $\lambda = 1$. The intensity of the total Poisson process for all crossovers of the k^{th}pedigree with m_k meioses equals m_k and the total intensity for the whole pedigree set equals $m_{tot} = m_1 + m_2 + \ldots + m_N$ (the total number of meioses).

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

In this subsection we will give a brief introduction to the concept of importance sampling following the presentation in Asmussen (1999).

Let $\underline{\mathrm{Z}} = \{\bar{Z}(x);\ x \in \Omega\}$ and $f(\underline{\mathrm{Z}}) = I(\bar{Z}_{\max} \geq T)$. Then

$$\alpha = \alpha(T) = E_0(f(\underline{\mathrm{Z}})) = \int f(\underline{\mathrm{Z}}) dP_0(\underline{\mathrm{Z}}), \tag{10}$$

where E_0 denotes expectation under the P_0 distribution of no linkage. This value may be Monte Carlo estimated without bias, using J simulations, as $\hat{\alpha} = \frac{1}{J}\sum_{i=1}^{J} f(\underline{Z}_i)$, where $\{\underline{Z}_i\}$ are i.i.d. copies of $\underline{Z}$ $\Big($cf. (8)$\Big)$. A possible improvement may in some situations be introduced by changing the probability measure from P_0 to $\tilde{P}$ and considering the formula

$$\alpha = \alpha(T) = E_0\Big(f(\underline{Z})\Big) = \tilde{E}\Big(L(\underline{Z})f(\underline{Z})\Big) \tag{11}$$

with $L(\underline{Z}) = dP_0(\underline{Z})/d\tilde{P}(\underline{Z})$ the likelihood ratio. Formula (11) requires that

$$f(\underline{Z})dP_0(\underline{Z}) > 0 \Longrightarrow d\tilde{P}(\underline{Z}) > 0, \tag{12}$$

i.e. that the support of $\tilde{P}$ is a least as large as that of fP_0. Hopefully, we get a variance reduction when estimating α by

$$\tilde{\alpha} = \frac{1}{J}\sum_{i=1}^{J} L(\underline{Z}_i)f(\underline{Z}_i), \tag{13}$$

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

A possible choice of $\tilde{P}$ is

$$d\tilde{P}(\underline{Z}) = \frac{f(\underline{Z})}{E_0[f(\underline{Z})]}dP_0(\underline{Z}) = \frac{f(\underline{Z})}{\alpha}dP_0(\underline{Z}), \tag{14}$$

which gives $L(\underline{Z}_i)f(\underline{Z}_i) \equiv \alpha$ for all i and hence $V(\tilde{\alpha}) = 0$. This choice of $\tilde{P}$ is optimal but not possible to use in practice, since $\alpha = E_0(f(\underline{Z}))$ is unknown. However, this property may be seen as a guidance for choosing $\tilde{P}$ as close to (14) 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 of getting large values of the random variable $f(\underline{Z})$ of interest.

For further details on importance sampling cf. e.g. Hammersley and Handscomb (1964), Kotz and Johnson (1982a) and Hesterberg (1995).

3 Importance Sampling in NPL Analysis for Perfect Marker Data

3.1 One Single δ Value

We will pick $\tilde{P}$ from a family $\{P_\delta\}_{\delta \geq 0}$ of probability measures with the null hypothesis distribution P_0 corresponding to $\delta = 0$. Start by selecting a locus

x_0 according to a density $p(x_0)$ on Ω. Then generate the linkage score $\bar{Z}(x_0)$ in (5) at position x_0 in a way (depending on δ) that is described below. Finally, generate $\underline{Z} \mid \bar{Z}(x_0)$ according to the H_0 distribution. That is, for all non-x_0 chromosomes we generate $\bar{Z}(\cdot)$ as in section 2.4 and at the x_0 chromosome we propagate, because of the Markov property, all components of the inheritance vectors independently to the right and left of x_0 according to Markov chains with intensity matrices (9) and initial states as determined by the inheritance vector components at x_0.

For notational convenience, let us assume that the score function S is standardized, i.e. $S(\cdot) = Z(\cdot)$ in (4). When generating $\bar{Z}(x_0)$, we assume perfect marker information for each family score in (3), $\bar{Z}_k(x_0) = S_k(v_k(x_0))$, where v_k is the inheritance vector and S_k the score function of the k^{th}pedigree respectively. We choose $v_k(x_0)$ according to

$$P_\delta(v_k(x_0) = w) \propto \exp(\delta_k S_k(w)), \tag{15}$$

where $\delta_k = \delta\gamma_k$. This will then imply, assuming $\delta > 0$, that inheritance vectors corresponding to large scores are more likely to be chosen.

Next we turn to the likelihood ratio. In appendix A it is shown that

$$L_\delta(\underline{Z}) = c(\delta)\Big(\int_\Omega \exp\Big(\delta\bar{Z}(x_0)\Big)p(x_0)dx_0\Big)^{-1} \tag{16}$$

where $c(\delta) = 2^{-m_{tot}}\prod_{k=1}^{N}\sum_{w\in Z_2^{m_k}}\exp\Big(\delta\gamma_k S_k(w)\Big)$ and $Z_2^{m_k}$ is the set of binary vectors of length m_k. The definition of the total NPL score is given in (5) and in (15)-(16) it is assumed that $\sum_{k=1}^{N}\gamma_k^2 = 1$.

The distribution (15) was introduced by Kong and Cox (1997) as an empirical likelihood. Then likelihood techniques could be employed to estimate δ at each locus as a way to perform linkage analysis. Here, we use x_0 as an artificial disease locus with δ measuring the strength of the genetic component. It is shown in appendix C that $\bar{Z}(x_0)$ has approximately a normal, $N(\delta, 1)$, distribution for large sample sizes. This gives us a natural interpretation of the parameter δ. It is approximately the average linkage score at the artificial disease locus.

With a proper choice of $\delta = \delta(T)$ we want P_δ to mimic the optimal probability measure given in section 2.5. Using (14) this measure $\tilde{P}$ equals $P_0(Z \mid Z_{\max} \geq T)$. We can think of x_0 as an approximation of $\tilde{x}_0$, an imagined disease locus under $\tilde{P}$, and (15) as an approximate distribution of $v_k(\cdot)$ at this locus. Moreover, x_0 and $\tilde{x}_0$ might be seen as approximations of $x_{\max}$ and

$\tilde{x}_{\max}$, i.e. the loci where the maximal NPL score under P_δ and $\tilde{P}$ respectively is attained. The rationale for this analogy is that a large expected NPL score exceeding T gives the impression of a large (false) disease locus, located in close vicinity of the corresponding maximal NPL score. The larger T is, the closer $x_{\max}$ is to x_0 and $\tilde{x}_{\max}$ to the (hypothesized) $\tilde{x}_0$.

Our main goal is to find a good estimate of $\alpha = \alpha(T)$ in (10)-(11). The value T is a member of the preselected set of interesting thresholds i.e. $T \in \{T_1, T_2, \ldots, T_R\}$. Note that estimation may be performed simultaneously, through simulation, for all values of T. Let $\underline{Z}_1^\delta, \underline{Z}_2^\delta, \ldots, \underline{Z}_J^\delta$ be i.i.d. random variables from P_δ and introduce the empirical measure

$$\hat{P}_\delta = \frac{1}{J} \sum_{i=1}^{J} \delta_{\underline{Z}_i^\delta} \tag{17}$$

where $\delta_{\underline{Z}^\delta}$ is a point mass at $\underline{Z}^\delta$. This gives us the expression

$$\begin{aligned} \hat{\alpha}_\delta &= \hat{\alpha}_\delta(T) \\ &= \int f(\underline{Z}) L_\delta(\underline{Z}) \hat{P}_\delta(\underline{Z}) \\ &= \frac{1}{J} \sum_{i=1}^{J} f(\underline{Z}_i^\delta) L_\delta(\underline{Z}_i^\delta). \end{aligned} \tag{18}$$

The first two moments for this estimator are expressed as $E(\hat{\alpha}_\delta) = \alpha$ and $Var(\hat{\alpha}_\delta) = \sigma_\delta^2$.

3.2 Weighting Estimates for Several δ Values

The choice of P_δ is crucial in order for $\hat{\alpha}_\delta$ to achieve variance reduction compared to $\hat{\alpha}_0$. We now consider weighted averages of various $\hat{\alpha}_\delta$ values whose weights need to be chosen properly.

Consider M different δ-values; $0 = \delta^1 < \delta^2 < \cdots < \delta^M$. Given weights $w_1, w_2, \ldots, w_M$ where $w_i \geq 0$ and $\sum_{i=1}^M w_i = 1$ we define

$$\hat{\alpha}_w = \sum_{i=1}^{M} w_i \hat{\alpha}_{\delta^i} \tag{19}$$

and then the obvious expressions of the expected value and variance of this estimator are

$$E(\hat{\alpha}_w) = \alpha \qquad \text{and} \qquad Var(\hat{\alpha}_w) = \sum_{i=1}^{M} w_i^2 \sigma_{\delta^i}^2 \tag{20}$$

respectively. To minimize variance we use Tukey's inequality (Kotz and Johnson, 1982b) to define the values of the weights as

$$w_i \propto \sigma_{\delta^i}^{-2}, \tag{21}$$

which means that we now need to be able to estimate σ_δ^2 in a proper way in order to find good and adequate weights. Notice that the optimal weights (21) depend on T whereas all $\{\hat{P}_{\delta^i}\}_{i=1}^M$ do not. Hence it is just the weighting in (19) and the function f in (18) that depend on the threshold.

Introduce

$$\beta_\delta(T) = P_\delta(\bar{Z}_{\max} \geq T) \tag{22}$$

for the probability that the maximum NPL score exceeds T under P_δ. We interpret $\beta_\delta(T)$ as the power of the test $\bar{Z}_{\max} \geq T$ given H_1 to detect the artificial disease locus at x_0 under P_δ. Moreover, define an estimator

$$\hat{\beta}_\delta(T) = \hat{P}_\delta(\bar{Z}_{\max} \geq T) = \frac{1}{J}\sum_{i=1}^{J} I(\bar{Z}_{\max,i}^\delta \geq T). \tag{23}$$

After some algebra one finds that

$$\alpha = \alpha(T) = \beta_\delta(T) E_\delta\Big(L(\underline{Z}) \mid \bar{Z}_{\max} \geq T\Big) \tag{24}$$

and

$$\sigma_\delta^2 = \frac{1}{J}\Big(\beta_\delta(T) Var_\delta\Big(L(\underline{Z}) \mid \bar{Z}_{\max} \geq T\Big) + \beta_\delta(T)\Big(1-\beta_\delta(T)\Big) E_\delta\Big(L(\underline{Z}) \mid \bar{Z}_{\max} \geq T\Big)^2\Big) \tag{25}$$

An estimator $\hat{\sigma}_\delta^2$ of σ_δ^2 is defined by replacing $\beta_\delta(T)$, Var_δ and E_δ in (25) by the corresponding empirical quantities $\hat{\beta}_\delta(T)$, $\hat{Var}_\delta$ and $\hat{E}_\delta$. One problem with this estimator is a tendency to be noisy for small or large values of δ (in relation to T) which with a large probability might cause the corresponding weight to be too large. In addition, $\hat{\alpha}_\delta$ will be skewed, with a heavier tail to the right, and with high probability underestimate α when δ is too small or too large.

To avoid these effects we define the truncation rule of the weights given in (21), for a given value of T,

$$w_i \propto \hat{\sigma}_{\delta^i}^{-2} \qquad \text{if} \qquad \epsilon_1 \leq \hat{\beta}_{\delta^i}(T) \leq \epsilon_2, \tag{26}$$

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 interval $[\epsilon_1, \epsilon_2]$ in (26) depends both on the sample size J and the threshold T. On one hand, we wish to choose $[\epsilon_1, \epsilon_2]$ small enough so that less reliable weights w_i and estimates of α_{δ^i} are weighted down. On the other hand, we wish to choose $[\epsilon_1, \epsilon_2]$ large enough to avoid loss of information by retaining all good $\hat{\alpha}_{\delta^i}$-estimates. For instance, $\epsilon_1 = C/J$ corresponds to a coefficient of variation of $\hat{\alpha}_{\delta^i}$ slightly larger than $1/\sqrt{C}$, cf. (24)-(25). When $\beta_\delta(T)$ is too large, the distribution of the likelihood ratio, $L(\underline{Z}) \mid \bar{Z}_{\max} \geq T$, will be skewed, with a heavier tail to the right, under P_δ. This implies that the second factor in (16) will be underestimated with high probability if J is too small or T is too large. The upper bound ϵ_2 should therefore be an increasing function of J and a decreasing function of T.

The procedure (26) gives large weights w_i to small i when T is small and to large i when T is large. This means that we essentially, for relatively small T, perform unweighted simulation (using P_0; $w_1 = 1$) and then with increasing T we successively include and remove other $\hat{\alpha}_{\delta^i}$ according to (26). With high probability new estimates $\hat{\alpha}_{\delta^i}$ are included in order $i = 2, 3, \ldots..$ and removed in order $i = 1, 2, \ldots$.

A simpler way of choosing w_i in (19), the so called quick method, is described in appendix D.

3.3 The Split-Merge Method

In principle, the importance sampling scheme works when Ω consists of several chromosomes. However, since marker data from different chromosomes is independent, it is natural to use this information and split the p-value estimation procedure into distinct estimates for each chromosome belonging to Ω and then finally merge this information into a joint genome-wide p-value. We call this method the split-merge method.

The split-merge formula for genome-wide p-value estimation is

$$\hat{\alpha}_w = \hat{\alpha}_w(T) = 1 - \prod_{k=1}^{C} (1 - \hat{\alpha}_{w,k}(T)), \tag{27}$$

where $\hat{\alpha}_{w,k}(T)$ corresponds to the weighted overall p-value estimate for the k^{th}chromosome and C is the number of included chromosomes.

4 Importance Sampling in NPL Analysis for Incomplete Marker Information

It is possible to generalize the importance sampling procedure to incomplete marker data. By combining (3) and (5) we see that the NPL score $\bar{Z}(\cdot)$ is a function of marker data $\underline{\mathrm{MD}} = (\mathrm{MD}_1, ..., \mathrm{MD}_N)$, where MD_k is the marker data for the k^{th}pedigree. In order to get a manageable expression for the likelihood ratio, we must define P_δ as a function of $\underline{\mathrm{MD}}$ rather than $\bar{Z}(\cdot)$ when marker data is incomplete. $\underline{\mathrm{MD}}$ is a function of inheritance vectors $\underline{\mathrm{v}} = \{v_k(x); x \in \Omega, k = 1, ..., N\}$ and founder genotypes at all marker loci for all pedigrees, $\underline{\mathrm{MD}}_{fnd}$. We simulate $\underline{\mathrm{v}}$ as described in Section 3.1 and appendix A. Then $\underline{\mathrm{MD}}_{fnd}$ is simulated independently of $\underline{\mathrm{v}}$, with a distribution not depending on δ. It is shown in appendix B that

$$\begin{aligned} L(\underline{\mathrm{MD}}) &= \frac{P_0(\underline{\mathrm{MD}})}{P_\delta(\underline{\mathrm{MD}})} \\ &= c(\delta) \left(\int_\Omega p(x_0) \prod_{k=1}^{N} \sum_{w \in Z_2^{m_k}} \exp\Big(\delta \gamma_k S_k(w)\Big) P(v_k(x_0) = w \mid \mathrm{MD}_k) dx_0 \right)^{-1}, \end{aligned} \tag{28}$$

where S_k is the score function for the k^{th}pedigree, S_k is standardized to have zero mean and unit variance one under H_0 (cf. $\big(4\big)$) and the weights satisfy $\sum_{k=1}^{N} \gamma_k^2 = 1$. The importance sampling estimate of α for one δ is

$$\hat{\alpha}_\delta(T) = \frac{1}{J} \sum_{i=1}^{J} L(\underline{\mathrm{MD}}^i) f(\underline{\mathrm{MD}}^i), \tag{29}$$

where $\{\underline{\mathrm{MD}}^i\}_{i=1}^{J}$ are i.i.d. drawn from P_δ and $f(\underline{\mathrm{MD}}) = I(\bar{Z}_{\max} \geq T)$. Estimates for several δ are combined in the same way as described in Section 3.2. The likelihood ratio (28) involves the conditional inheritance distributions $P(v_k(x) \mid \mathrm{MD}_k)$ 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 (3) and hence has to be computed for all pedigrees at all loci in order to define $\bar{Z}_{\max}$. Therefore, the *additional* computational burden to evaluate the likelihood ratio is relatively small.

5 Results

Throughout this section we let $p(x_0)$ have a uniform distribution over Ω. We use equal weighting of pedigrees, i.e. $\gamma_k = 1/\sqrt{N}$ for $k = 1, \ldots, N$ and homogenous pedigree sets, i.e. all pedigrees have the same structure, chosen from figure 1. In most of our simulations we use $N = 60$. Using this and $\{\delta_k^i\}_{i=1}^M = \{0, 0.05, 0.1, \ldots, 0.7\}$ correspond to the following grid of δ-values,

$$\{\delta^i\}_{i=1}^M = \{0, 0.3873, 0.7746, \ldots, 5.4222\}. \tag{30}$$

We compare, for perfect marker data, the weighted simulations to traditional unweighted simulations and to an approximation formula based on extreme value theory for stochastic processes Ängquist and Hössjer (2003).

5.1 Effect of Choosing Tuning Constants

Here we define the artificial genome to consist of one single chromosome with a corresponding genetic length equal to $l = 3.0$ Morgans.

Firstly, we consider the assumption of normally distributed NPL scores at the random position x_0. Here we use the structure of pedigree 1. As shown in appendix C, $\bar{Z}(x_0)$ may be approximated by a $N(\delta, 1)$ distribution. In figure 2 this is displayed for $N = 60$ and $N = 5$ respectively, for $\delta_k = \delta = 0$ and $\delta_k = \delta/\sqrt{N} = 0.7$. As expected the approximation is slightly more accurate with larger N and smaller δ_k, but the differences are very modest for this pedigree.

One may note that the importance sampling estimator has *tuning constants*. For the estimator (19) we have $\{\delta^i\}_{i=1}^M$, J, ϵ_1 and ϵ_2. For the quick method (50) we replace the latter two quantities with a single ϵ. For the remaining part of this subsection we use pedigree number three and further details about the simulations are given in the figure captions.

Secondly, we investigate an example of the successively inclusion and exclusion of the weights $w_1, \ldots, w_M$, cf. figure 3. The results seem to be consistent with the previous discussion. One may notice that a common number of contemporary active δ-values during the analyses, in this setting, is 2-5 (less for very small and very large Ts). This corresponds to a range of δ roughly between 1 and 2, cf. (30).

Thirdly, we investigate the importance of the total number J of simulations performed, see figure 4. The importance sampling estimator seems to

be very robust, even for surprisingly small p-values, with respect to the tuning constants involved. A general interpretation may be that for small J it is advantageous to use a large ϵ_2 to keep enough information and for a large J it is possible to decrease ϵ_2 in order to avoid underestimation of $\alpha(T)$ for large thresholds. For the quick method (cf. figure 5) we choose $\epsilon = 0.0228$ or $\epsilon = 0.00135$ so that $\hat{\delta}$ in (49) equals $T-2$ and $T-3$ respectively.

Fourthly, we perform simulations with respect to different $[\epsilon_1, \epsilon_2]$-combinations, cf. figures 6-7. The results do not seem to be very sensitive for the specification of this acceptance interval, i.e. the method is robust in this respect.

5.2 Four Examples of Full-Scale Autosomal Investigations

Here we used all four pedigree sets to further test the importance sampling simulation procedure in a full autosomal setting.

In this case we perform a split-merge analysis (section 3.3) and a genome region consisting of all the 22 human autosomes, cf. figure 8. The procedure is successful i.e. we are able to get good estimates for p-values corresponding to a wide range of thresholds. 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. Using $J = 3000$ and $M = 15$ in the former case leads to accurate estimates of p-values of magnitude less than 10^{-10} to be compared to approximately 10^{-5}, when $J = 50000$, in the latter case.

5.3 Computational Complexity

We investigate the relation of computational complexity between unweighted simulation and a split-merge importance sampling procedure. We use pedigree 1, $N = 60$, $C = 5$, $l = 2$ Morgans, $R = 41$ thresholds, $J \in \{10, 50, 100, 500, 1000\}$ and $M \in \{1, 3, 5, 10, 15\}$. For the unweighted simulation the computation time $t \approx C_1 J$, whereas for the weighted simulation $t \approx C_2 M J$ and the ratio of the constants C_2/C_1 is approximately equal to 3. This means that importance sampling is faster if J is reduced by a factor $3M$ or more compared to traditional simulation. If we are to compute $\alpha(T)$ for a single T, we may use the simple method with only one δ $\Big($cf. (49)$\Big)$, i.e. $M = 1$. In this case, the computational complexity is often much smaller for importance sampling.

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 constant number of simulations, accurately estimate quantitatively very small p-values. Less simulations are needed to, given a constant accuracy, adequately estimate a specific small p-value. 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.

Moreover, we have generally assumed perfect data but have also (in section 4) generalized the method to incomplete marker information. One reason for mainly assuming perfect marker data is that it is then possible to compare the simulation results to the approximation formula given in Ängquist and Hössjer (2003). Moreover, to analyze the performance of this method we found it appropriate to start with a simpler case.

It is important to notice that though the number of simulations needed to get satisfactory results decreases when using this method, the time cost for each simulation obviously increases. This forces the implementation of simulation algorithms to become an important factor, but this does not seem to be an overwhelming problem, cf. section 5.3.

7 Acknowledgements

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.

A Likelihood Ratio for Perfect Marker Data

For simplicity, we consider the case $\Omega = [0, l]$ of one single chromosome of length l Morgans. In order to derive the likelihood ratio $L(\underline{\mathrm{Z}}) = dP_0(\underline{\mathrm{Z}})/dP_\delta(\underline{\mathrm{Z}})$, it is convenient to first introduce $\underline{\mathrm{v}}(x_0) = \Big(v_1(x_0), ..., v_N(x_0)\Big)$, $\underline{v}_k = \{v_k(x)\ ;\ x \in \Omega\}$ and $\underline{v} = \{\underline{v}_1, \ldots, \underline{v}_N\}$ for the inheritance processes. Then

$$dP_\delta(\underline{\mathrm{v}} \mid x_0) = P_\delta\Big(\underline{\mathrm{v}}(x_0)\Big) dP_\delta\Big(\underline{\mathrm{v}} \mid \underline{\mathrm{v}}(x_0)\Big). \tag{31}$$

Let us concentrate on the expression on the right-hand side in (31). We find that

i) The first factor may be expanded as

$$\begin{aligned} P_\delta\Big(\underline{\mathrm{v}}(x_0)\Big) &= \prod_{k=1}^{N} P_\delta\Big(v_k(x_0)\Big) = \prod_{k=1}^{N} \left(\frac{\exp\Big(\delta_k S_k\Big(v_k(x_0)\Big)\Big)}{c_k(\delta_k)} \right) \\ &= 2^{-m_{tot}} c(\delta)^{-1} \exp\Big(\delta \bar{Z}(x_0)\Big), \end{aligned} \tag{32}$$

where $c_k(\delta_k) = \sum_w \exp\Big(\delta_k S_k(w)\Big)$, and the normalizing factor $c(\delta)$ is defined below (16). We used independence of pedigrees and in the last step that $\bar{Z}_k(x) = S_k\Big(v_k(x)\Big)$ and $\bar{Z}(x) = \sum_{k=1}^{N} \gamma_k \bar{Z}_k(x)$, which coincides with (3) and (5) if we assume $\sum_{k=1}^{N} \gamma_k^2 = 1$.

ii) For the second factor we first consider the case $m_{tot} = 1$. Let $K \geq 0$ be the total number of crossovers, with positions $0 \leq T_1 \leq \ldots \leq T_K \leq l$. Define $j = j(x_0)$ by $T_j \leq x_0$ and $T_{j+1} > x_0$. (To make j well defined, we put $T_0 = -\infty$ and $T_{K+1} = \infty$.) Given $\underline{\mathrm{v}}(x_0)$, the distribution of $\underline{\mathrm{v}}$ is fully specified by the fact that the set of all crossovers evolve as two independent Poisson processes with intensity 1 to the right and left of x_0 respectively. Therefore

$$\begin{aligned} dP_\delta\Big(\underline{\mathrm{v}} \mid \underline{\mathrm{v}}(x_0)\Big) &= dP_\delta\Big(\underline{\mathrm{v}}(x_0-) \mid \underline{\mathrm{v}}(x_0)\Big) dP_\delta\Big(\underline{\mathrm{v}}(x_0+) \mid \underline{\mathrm{v}}(x_0)\Big) \\ &= p(T_1, ..., T_j) p(T_{j+1}, \ldots, T_K) d\underline{\mathrm{v}} \\ &= \exp(-x_0) \exp\Big(-(l - x_0)\Big) d\underline{\mathrm{v}} \\ &= \exp(-l) d\underline{\mathrm{v}}, \end{aligned} \tag{33}$$

where $\underline{\mathrm{v}}(x_0-) = \{\underline{\mathrm{v}}(x);\ 0 \leq x < x_0\}$, $\underline{\mathrm{v}}(x_0+) = \{\underline{\mathrm{v}}(x);\ x_0 < x \leq l\}$ and $p(T_1, ..., T_j)$ is the joint density of $T_1, ..., T_j$. For a general number of meioses

we simply multiply m_{tot} factors of the kind (33) to arrive at

$$dP_\delta(\underline{\mathrm{v}} \mid \underline{\mathrm{v}}(x_0)) = \exp(-m_{tot}l)d\underline{\mathrm{v}}. \tag{34}$$

Now $\underline{\mathrm{Z}} = \underline{\mathrm{Z}}(\underline{\mathrm{v}})$ is a function of $\underline{\mathrm{v}}$. Let $A = A(\underline{\mathrm{Z}})$ be defined by $A = \{\underline{\mathrm{v}}\,;\, \underline{\mathrm{Z}}(\underline{\mathrm{v}}) = \underline{\mathrm{Z}}\}$. Then, combining (31), (32) and (34) we derive

$$\begin{aligned} dP_\delta(\underline{\mathrm{Z}}) &= \int_0^l \int_A dP_\delta(\underline{\mathrm{v}} \mid x_0)p(x_0)dx_0 \\ &= \int_0^l P_\delta\Big(\underline{\mathrm{v}}(x_0)\Big) \int_A dP_\delta\Big(\underline{\mathrm{v}} \mid \underline{\mathrm{v}}(x_0)\Big)p(x_0)dx_0 \\ &= 2^{-m_{tot}}c(\delta)^{-1}\exp(-m_{tot}l)\int_A d\underline{\mathrm{v}} \int_0^l \exp\Big(\delta\bar{Z}(x_0)\Big)p(x_0)dx_0. \end{aligned} \tag{35}$$

Finally, formula (16) follows by applying (35) twice, once for δ and once for 0, and then computing the likelihood ratio.

B Likelihood Ratio under Incomplete Marker Information

As in appendix A, we consider, for simplicity, the case of one chromosome $\Omega = [0, l]$. The marker data is a function of $\underline{\mathrm{v}}$ and the founder alleles at all marker loci for all pedigrees. We write $\underline{\mathrm{MD}} \sim \underline{\mathrm{v}}$ if marker data is consistent with $\underline{\mathrm{v}}$. That is, at all marker loci segregation of marker alleles is consistent with $\underline{\mathrm{v}}$ at the corresponding loci and for all pedigrees. The requirement $\underline{\mathrm{MD}} \sim \underline{\mathrm{v}}$ is less stringent if markers have a low heterozygosity since then more segregation patterns are possible. Define $B = B(\underline{\mathrm{MD}})$ by $B = \{\underline{\mathrm{v}}\,;\, \underline{\mathrm{MD}} \sim \underline{\mathrm{v}}\}$. Then

$$P_\delta(\underline{\mathrm{MD}}) = \int_B P(\underline{\mathrm{MD}} \mid \underline{\mathrm{v}})dP_\delta(\underline{\mathrm{v}}). \tag{36}$$

By Bayes' rule, the conditional inheritance distribution of $\underline{\mathrm{v}}$ given markers can be written as

$$dP(\underline{\mathrm{v}} \mid \underline{\mathrm{MD}}) = \frac{P(\underline{\mathrm{MD}} \mid \underline{\mathrm{v}})dP_0(\underline{\mathrm{v}})}{P_0(\underline{\mathrm{MD}})}. \tag{37}$$

From appendix A it follows that

$$dP_\delta(\underline{\mathrm{v}}) = c(\delta)^{-1}\int_0^l p(x_0)\prod_{k=1}^N \exp\Big(\delta\gamma_k S_k\Big(v_k(x_0)\Big)\Big)dx_0 dP_0(\underline{\mathrm{v}}). \tag{38}$$

Inserting these two equations into (36) and changing the order of integration we get

$$P_\delta(\underline{\mathrm{MD}}) = P_0(\underline{\mathrm{MD}})c(\delta)^{-1}\int_0^l p(x_0)\int_B \prod_{k=1}^N \exp\Big(\delta\gamma_k S_k\big(v_k(x_0)\big)\Big)dP(\underline{\mathrm{v}}\mid\underline{\mathrm{MD}})dx_0. \tag{39}$$

Because of independence of marker data for different pedigrees

$$dP(\underline{\mathrm{v}}\mid\underline{\mathrm{MD}}) = \prod_{k=1}^N dP(\underline{\mathrm{v}}_k\mid \mathrm{MD}_k). \tag{40}$$

Therefore, the inner integral in (39) may be written as

$$\prod_{k=1}^N\sum_w \exp\Big(\delta\gamma_k S_k(w)\Big)P(v_k(x)=w\mid \mathrm{MD}_k). \tag{41}$$

Combining (39), (41) and taking the likelihood ratio we obtain (28).

C Asymptotic Distribution of $\bar{Z}(x_0)$

We begin with deriving an approximation of the expected NPL score at x_0 under the probability measure P_δ when marker data is perfect. First consider one single family, the k^{th}pedigree. Put $p_{\delta_k}(z) = P_\delta(\bar{Z}_k(x_0)=z)$ when marker data is perfect $\big($cf. (15)$\big)$. Then

$$\begin{aligned} E_\delta(\bar{Z}_k(x_0)) &= \sum_z z\ p_{\delta_k}(z) \\ &= \frac{\sum_z z\ p_0(z)\exp(\delta_k z)}{\sum_z p_0(z)\exp(\delta_k z)} \\ &= \frac{M'(\delta_k)}{M(\delta_k)}, \end{aligned} \tag{42}$$

where $M(\delta_k) = E_0\Big(\exp\big(\delta_k\bar{Z}_k(x_0)\big)\Big)$ is the moment generating function of $\bar{Z}_k(x_0)$ under the null hypothesis H_0. Now

$$\begin{aligned} M(\delta_k) &= 1+\delta_k E_0\big(\bar{Z}_k(x_0)\big)+\frac{\delta_k^2}{2}E_0\big(\bar{Z}_k(x_0)^2\big)+o(\delta_k^2) \\ &= 1+\frac{\delta_k^2}{2}+o(\delta_k^2), \end{aligned} \tag{43}$$

which gives

$$\frac{M'(\delta_k)}{M(\delta_k)} = \delta_k + o(\delta_k). \tag{44}$$

as $\delta_k \to 0$.

For the total linkage score, using the definitions $\bar{Z}(x) = \sum_{k=1}^{N} \gamma_k \bar{Z}_k(x)$, $\sum_{k=1}^{N} \gamma_k^2 = 1$ and $\delta_k = \gamma_k \delta$ we get

$$\begin{aligned} E_\delta\Big(\bar{Z}(x_0)\Big) &= \sum_{k=1}^{N} \gamma_k E_\delta\Big(\bar{Z}_k(x_0)\Big) \\ &= \sum_{k=1}^{N} \gamma_k\Big(\delta_k + o(\delta_k)\Big) \\ &= \delta + o(1). \end{aligned} \tag{45}$$

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

Further, a similar analysis shows that

$$V_\delta\Big(\bar{Z}(x_0)\Big) = 1 + o(1) \tag{46}$$

as $N \to \infty$. Then a Central Limit Theorem argument implies, under mild regularity conditions on the set of pedigree structures, that

$$\bar{Z}(x_0) \xrightarrow{\mathcal{D}} N(\delta, 1) \tag{47}$$

as $N \to \infty$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

D The Quick Method

It is possible to use the assumption of approximately normally distributed NPL-scores at x_0 (cf. appendix C) to choose weights w_i in (19) in a simpler way. This method avoids the use of (21)-(26) and will henceforth be referred to as the quick method.

The score $\bar{Z}(x_0)$ is approximately $N(\delta, 1)$-distributed. Then

$$\begin{aligned} \beta_\delta(T) &= P_\delta(\bar{Z}_{\max} \ge T) \ge P_\delta(Z(x_0) \ge T) \\ &\approx 1 - \Phi\left(\frac{T - E_\delta\Big(\bar{Z}(x_0)\Big)}{\sqrt{V_\delta\Big(\bar{Z}(x_0)\Big)}}\right) \approx 1 - \Phi(T - \delta). \end{aligned} \tag{48}$$

In (26) we used an acceptance region $[\epsilon_1, \epsilon_2]$ to define the range of tolerable $\beta_\delta(T)$-values. Here we use one single ϵ and derive one single estimate $\hat{\delta} = \hat{\delta}(T)$

of the value $\delta = \delta(T)$ that solves $\beta_\delta(T) = \epsilon$. Motivated by (48) we define $\hat{\delta}$ by

$$1 - \Phi(T - \hat{\delta}) = \epsilon \Rightarrow \hat{\delta} = T - \Phi^{-1}(1 - \epsilon). \tag{49}$$

As an example, if we set $\epsilon = 0.0228$ we get $\hat{\delta} = T - 2$. In practice one may use a linear combination of the estimates corresponding to the two δ^i surrounding $\hat{\delta}$. Consequently

$$\hat{\alpha}_w = \hat{\alpha}_{w,\delta}(T) = w_1 \cdot \hat{\alpha}_{\delta^i}(T) + w_2 \cdot \hat{\alpha}_{\delta^{i+1}}(T), \tag{50}$$

where $w_1 = \left(\frac{\delta^{i+1} - \delta}{\delta^{i+1} - \delta^i}\right)$, $w_2 = \left(\frac{\delta - \delta^i}{\delta^{i+1} - \delta^i}\right)$ and $\delta^i \leq \hat{\delta} < \delta^{i+1}$.

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.

Abney, M., Ober, C. and McPeek, M.S. (2002). Quantitiative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: Fsting serum-insulin level in the Hutterites. *American Journal of Human Genetics, 70*, 920-934.

Ängquist, L. (2001). *Conditional two-locus NPL-analyses: Theory and applications* (Master's Thesis No. 2001:E22). Department of Mathematical Statistics, Lund University.

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

Asmussen, S. (1999). *Stochastic simulation with a view towards stochastic processes.* (Lecture Notes: www.maphysto.dk)

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. *Proceeding 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.

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.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods*, Wiley, New York.

Hesterberg, R. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics, 37(2)*, 185-194.

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

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 Johnstone, N.L. (1982a). Importance sampling. In S. Kotz and N.L. Johnson (Eds.), *Encyclopedia in statistical sciences* Vol. 4, p. 25. John Wiley & Sons.

Kotz, S. and Johnstone, N.L. (1982b). Tukey's inequality for optimal weights. In S. Kotz and N.L. Johnson (Eds.), *Encyclopedia in statistical sciences* Vol. 9, pp. 361-362. John Wiley & Sons.

Kryglyak, 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 quantitiative traits using RFLP maps. *Genetics, 121*, 185-199.

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. *Proceeding 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.). 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.

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.

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.

List of Figures

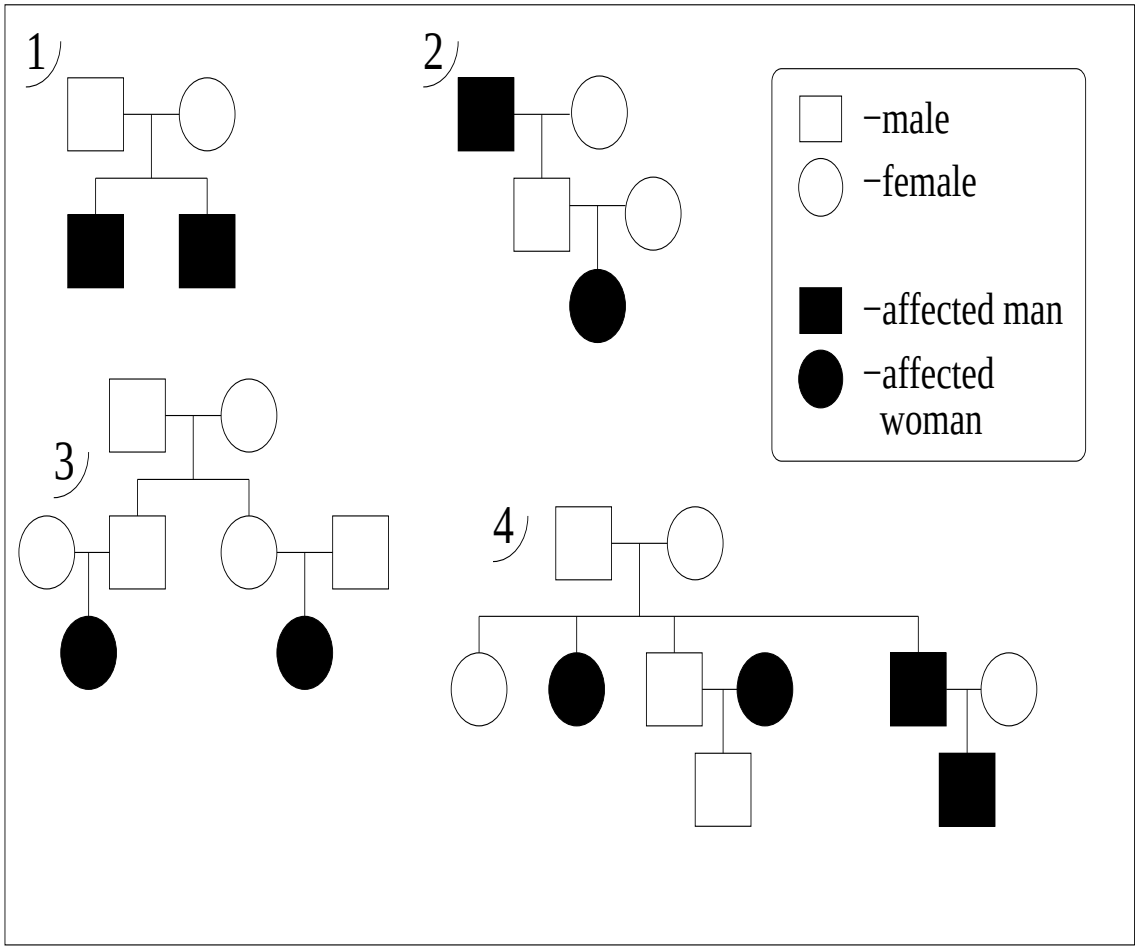

Figure 1: Four pedigrees of different structure.

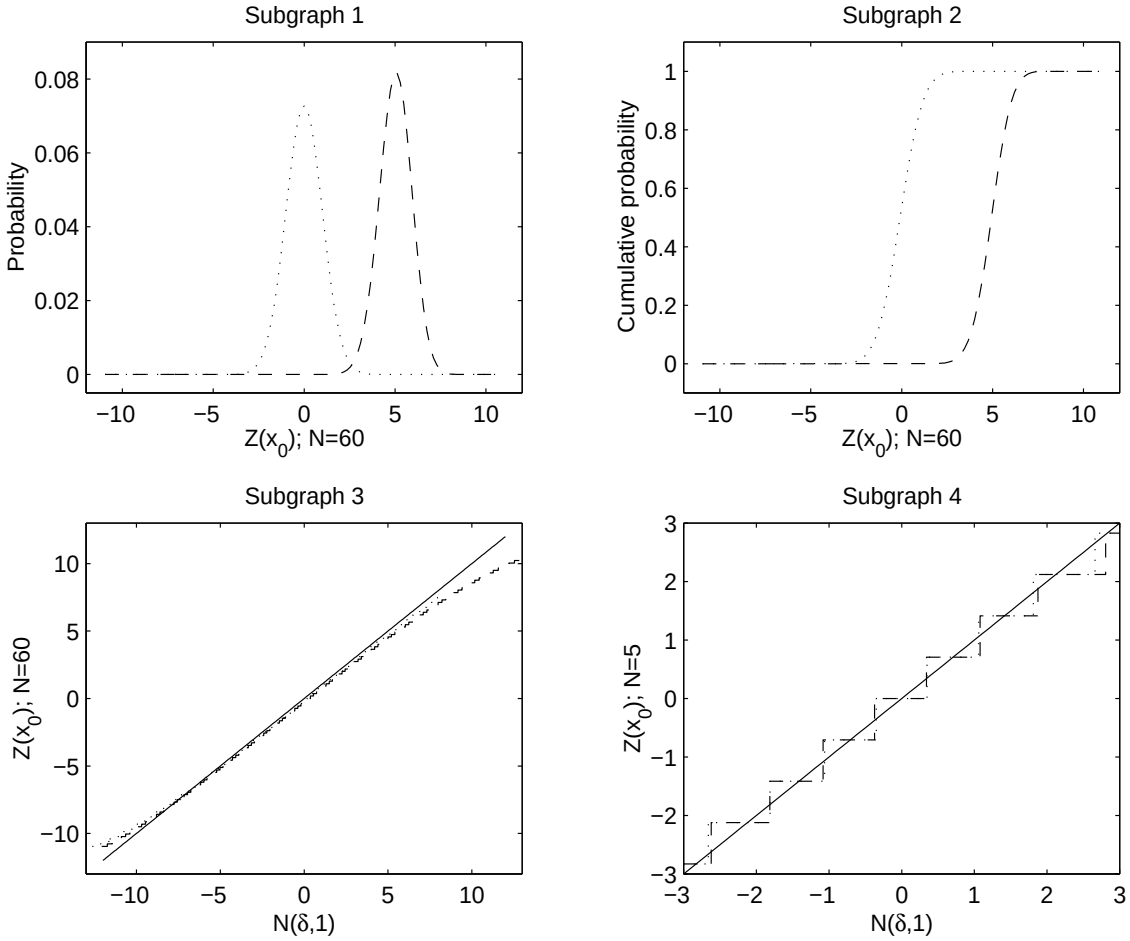

Figure 2: Pedigree 1, $\delta_k = \delta = 0$ $(\cdots)$ and $\delta_k = \delta/\sqrt{N} = 0.7$ $(--)$. 1: Probability distribution of $\bar{Z}(x_0)$. 2: Distribution function for $\bar{Z}(x_0)$. 3 and 4: Quantiles of $\bar{Z}(x_0)$ plotted against $N(\delta, 1)$–quantiles for N=60 (3) and N=5 (4). Plots compared to straight line $y = x$ $(-)$.

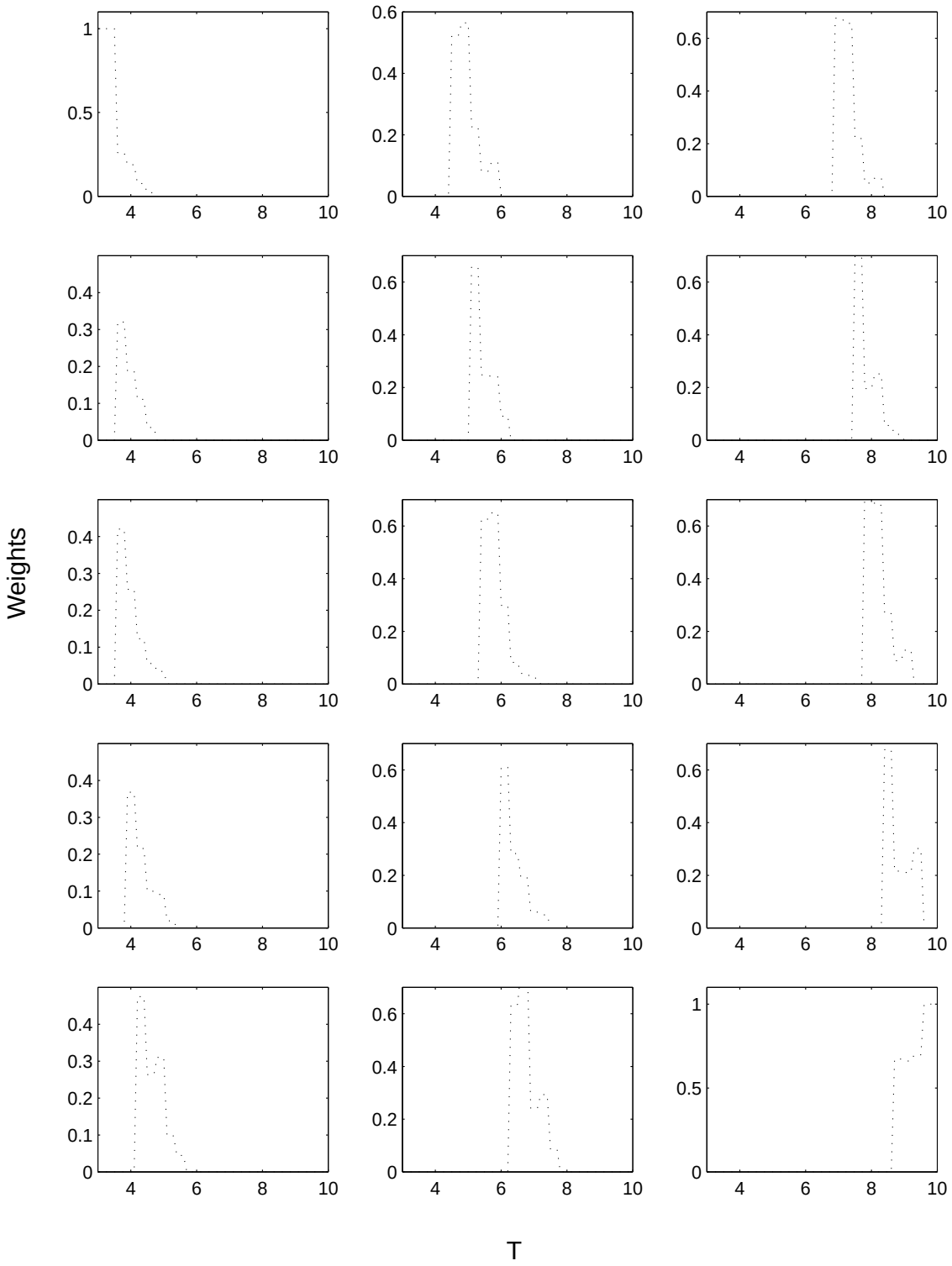

Figure 3: An example of the dependencies between the different weights and the thresholds. Pedigree 3. Parameters: J=10000, N=60, $[\delta^1, \delta^2, \ldots, \delta^{15}]$=[0.0000, 0.3873,..., 5.4222] and $[\epsilon_1, \epsilon_2]$=[0.001, 0.05].

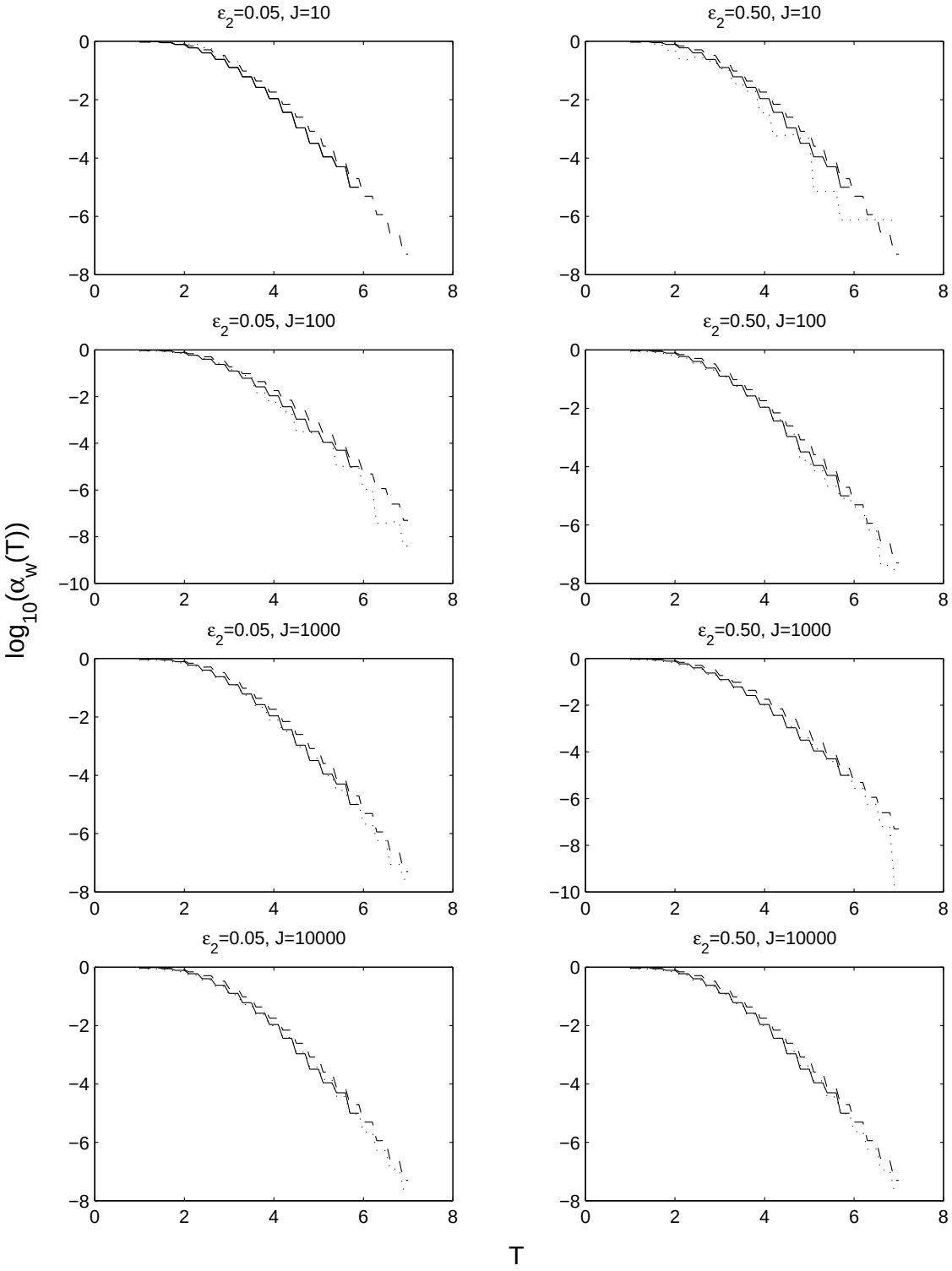

Figure 4: An example of the impact of the number of simulations J. Importance sampling ($\cdots$) vs. traditional unweighted simulation with J=100000 ($-$) and approximation formula ($--$). Pedigree 3. Parameters: N=60, $\{\delta^i\}_{i=1}^{M}$=[0.0000, 0.3873,..., 3.873], $\{T_i\}_{i=1}^{R}$=[1.0, 1.1,..., 7.0] and ϵ_1=0.001.

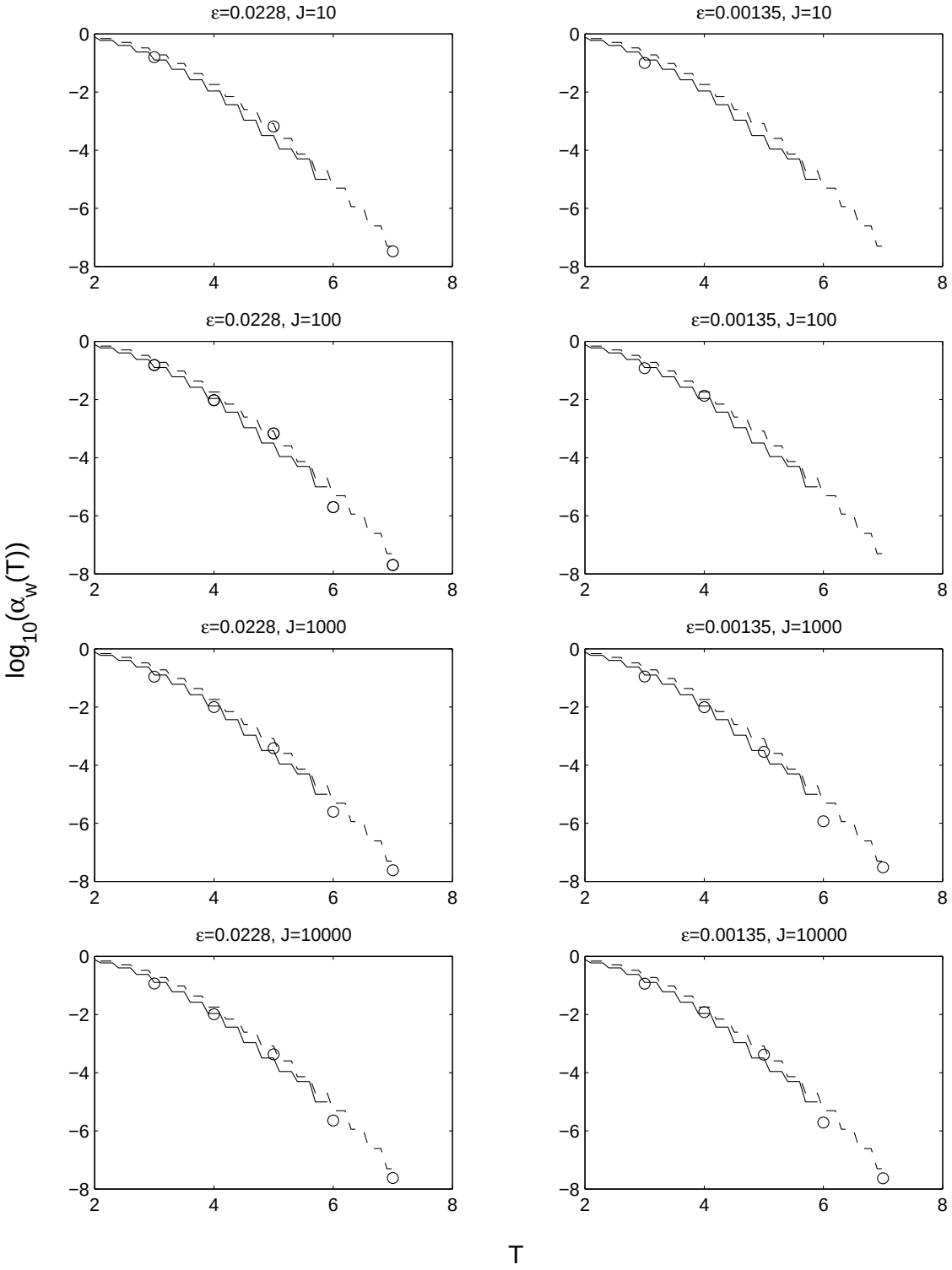

Figure 5: Simulation results for the quick method - an example. Importance sampling (o) vs. traditional unweighted simulation with J=100000 (–) and approximation formula (––). Pedigree 3, N=60, $\{T_i\}_{i=1}^{R}$=[3.0, 4.0, 5.0, 6.0, 7.0], $\hat{\delta}(T) = (T-2)$ (left) and $\hat{\delta}(T) = (T-3)$ (right). For missing data (i.e. circles) $\bar{Z}^i_{\max} < T$ for all simulations; $1 \leq i \leq J$.

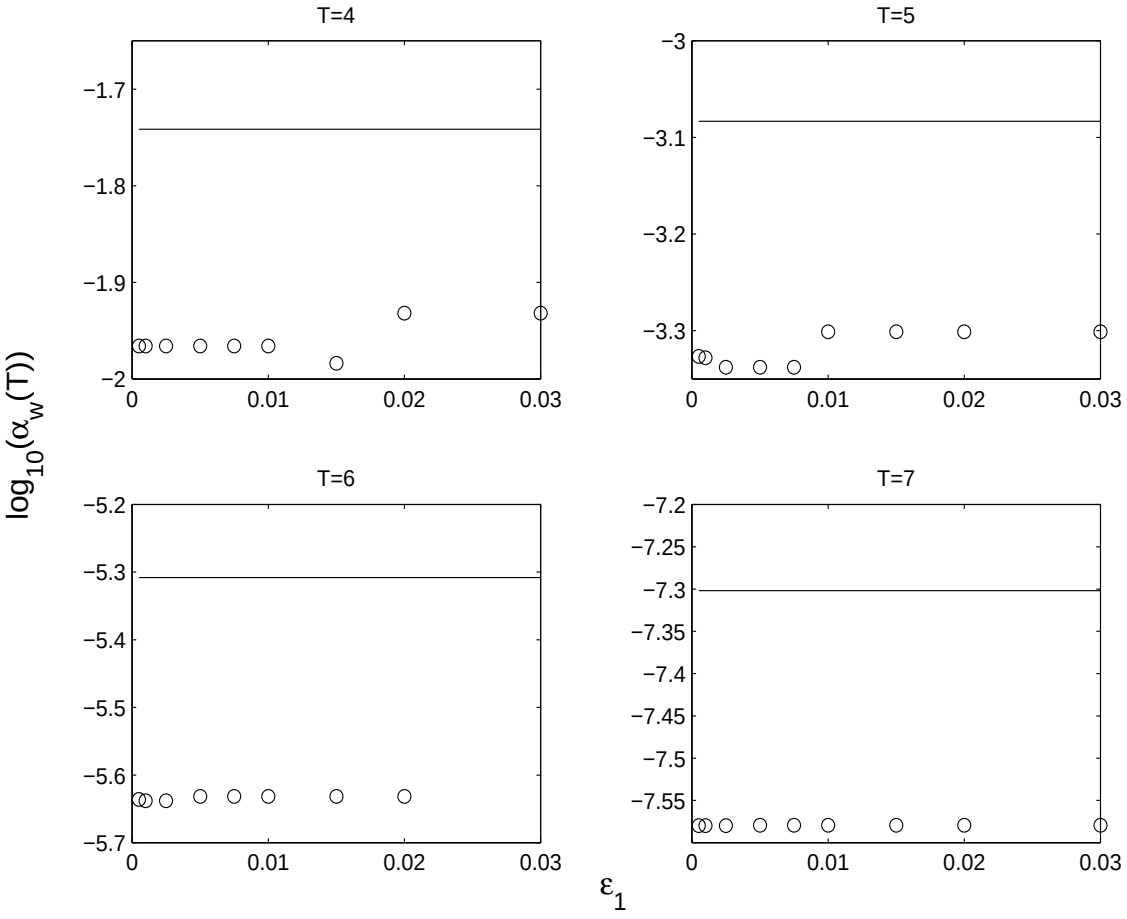

Figure 6: Sensitivity of importance sampling procedure with respect to variations in ϵ_1. Pedigree 3, ϵ_2=0.05 throughout, J=10000, N=60, $\{\delta^i\}_{i=1}^M$=[0.0000, 0.7746,..., 5.4222], $\{T_i\}_{i=1}^R$=[4.0, 5.0, 6.0, 7.0] and $\epsilon_1 \in$ [0.5, 1.0, 2.5, 5.0, 7.5, 10, 15, 20, 30]$\cdot 10^{-3}$. Importance sampling (o) vs. approximation formula (–).

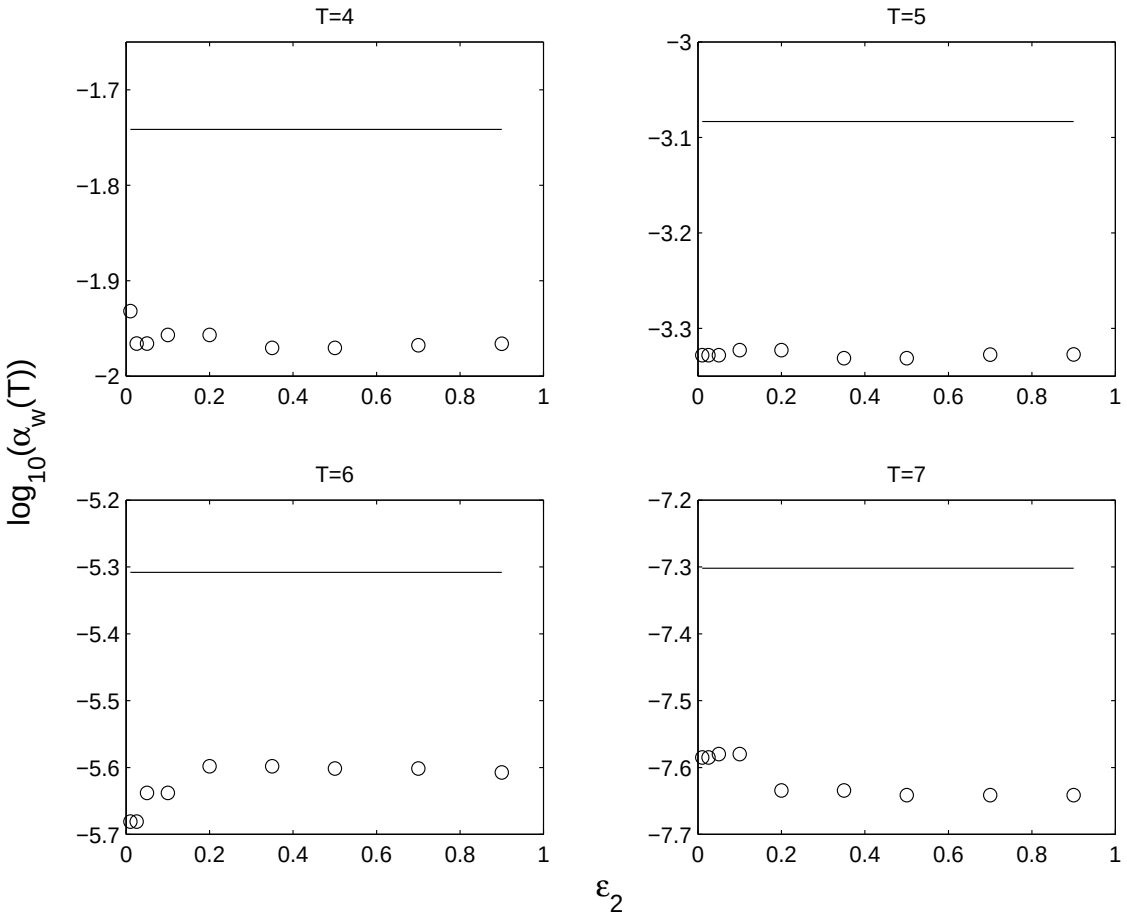

Figure 7: Sensitivity of importance sampling procedure with respect to variations in ϵ_2. Pedigree 3, ϵ_1=0.001 throughout, J=10000, N=60, $\{\delta^i\}_{i=1}^M$=[0.0000, 0.7746,..., 5.4222], $\{T_i\}_{i=1}^R$=[4.0, 5.0, 6.0, 7.0] and $\epsilon_2 \in$ [1.0, 2.5, 5.0, 10, 20, 35, 50, 70, 90]$\cdot 10^{-2}$. Importance sampling (o) vs. approximation formula (–).

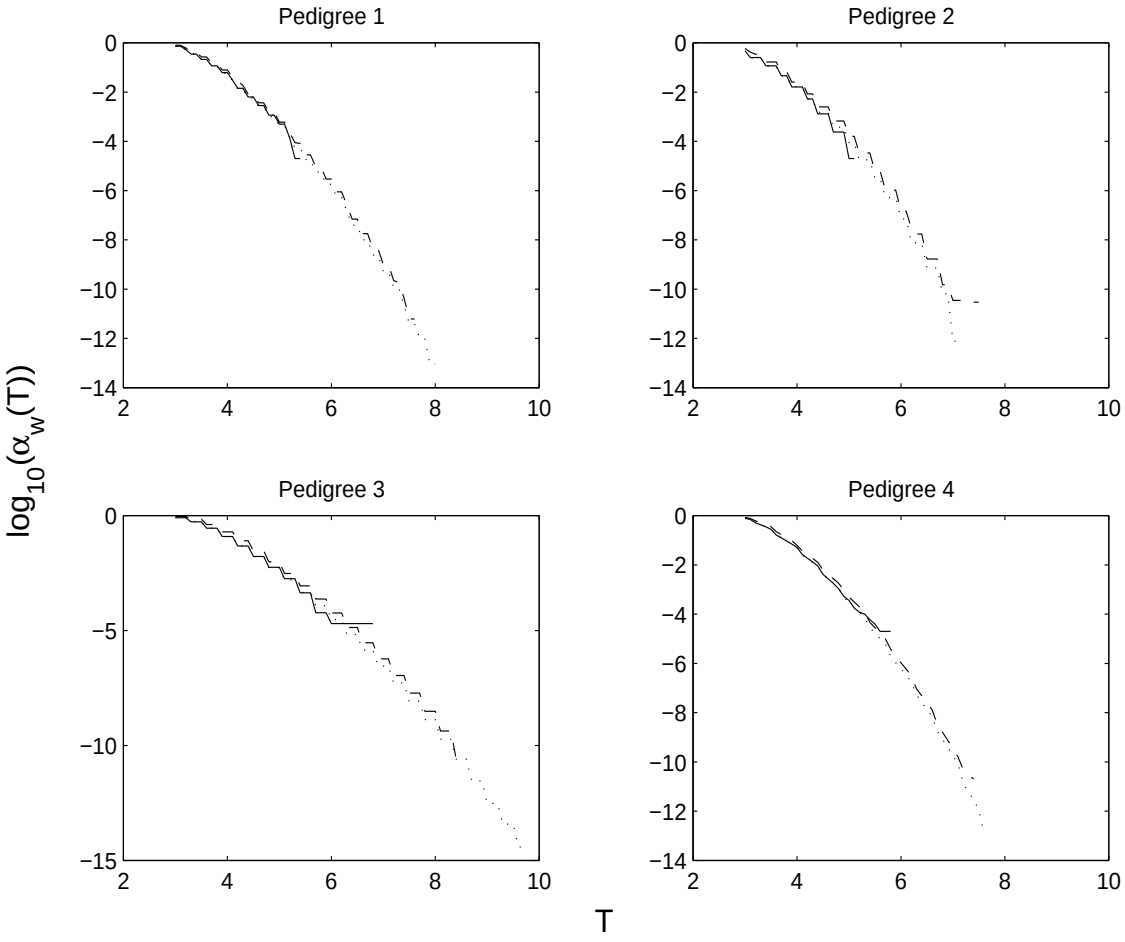

Figure 8: Significance comparisons between importance sampling with J=3000 ($\cdots$), unweighted simulation with J=50000 (–) and an approximation formula ($--$). The simulation is performed over all the 22 autosomes with a total chromosomal genetic length of 35.75 Morgans (Collins et al., 1996). Additional parameters: N=60, $\{\delta^i\}_{i=1}^M$=[0.0000, 0.3873,..., 5.4222] and $[\epsilon_1, \epsilon_2]$=[0.001, 0.05].