

Importance Sampling for Stochastic Processes

-

with Applications to Linkage Analysis

Ola Hössjer
Lars Ängquist

Linkage Analysis

- **Data:** A number of families with a certain disease and DNA marker data from (some of the) family members.
- **Goal:** 1) Test presence of and 2) Give confidence interval for location of disease gene on a certain region of the genome (e.g. chromosome).
- **Idea:** 'Interesting regions' along the chromosome has highest correlation (linkage) between inheritance of DNA and disease due to presence of crossovers (Morgan, 1911, Sturtevant, 1913).
- **Output:** A stochastic process

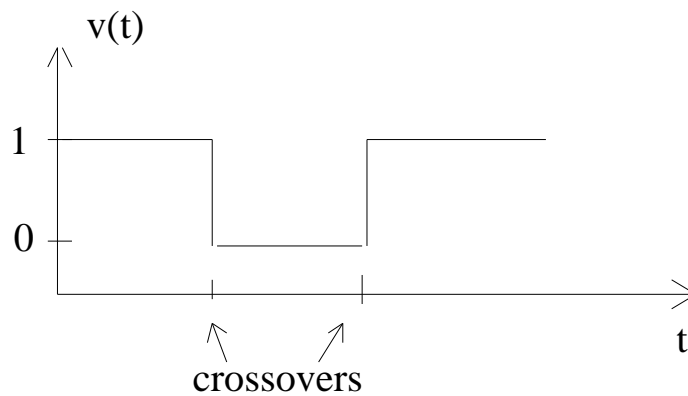
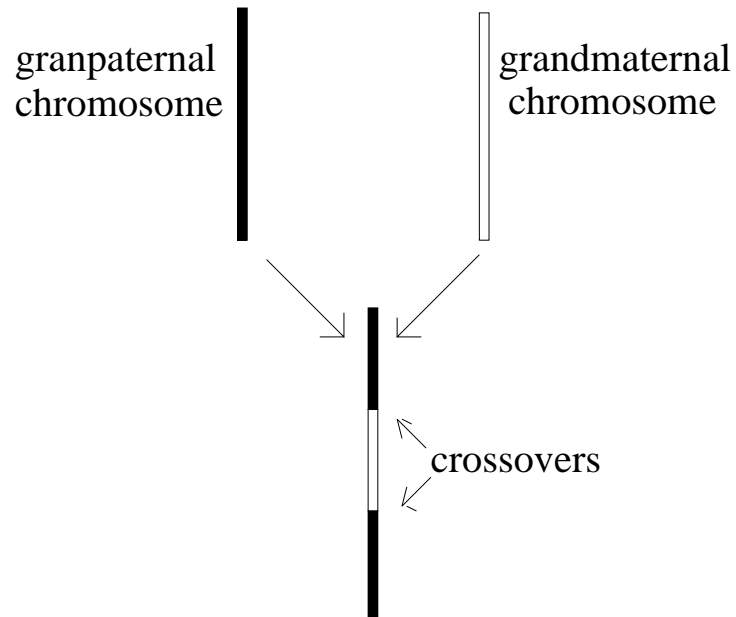
$$Z = \{Z(x), 0 \leq x \leq l\},$$

is defined (for each chromosome) and peaks of Z correspond to interesting regions.

- **Main question:** Is the highest peak of Z significant?

Crossovers

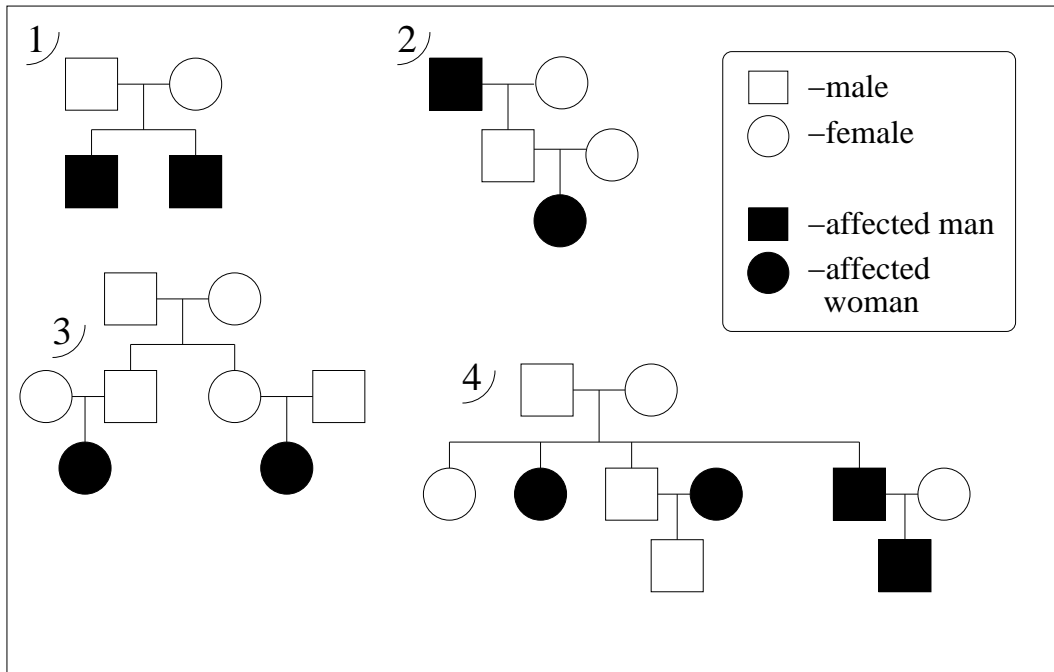
During *meiosis* (formation of germ cells), a number of *crossovers* occur:



$$E(\# \text{crossovers}) / \text{centiMorgan} = 0.01$$

Average chr length = 150 centiMorgan (cM).

Example Pedigrees



n = number of individuals
 f = number of founders
 m = $2(n - f)$ = number of meioses

Pedigree	n	f	m
1	4	2	4
2	5	3	4
3	8	4	8
4	10	4	12

Inheritance vector

For a pedigree with m meioses, define inheritance vector at position x (cM), as

$$\mathbf{v}(x) = (v_1(x), \dots, v_m(x)),$$

where

$$v_j(x) = 0 \quad \Rightarrow \text{grandmaternal transmission}$$

$$v_j(x) = 1 \quad \Rightarrow \text{grandpaternal transmission}$$

for j^{th} meiosis at position x (Donnelly, 1983).

Each member of pedigree has two copies of a gene (allele), one from the father and one from the mother. $\mathbf{v}(x)$ determines how alleles are segregated in pedigree at position x .

Mendel's law of segregation (Mendel, 1865) and Haldane's model for crossovers (Haldane, 1919) imply that $v_j(\cdot)$ evolve as m independent Markov processes in continuous 'time' with intensity matrix

$$\begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix}.$$

Nonparametric Linkage Analysis

Given chromosome of length l cM, we wish to test

H_0 : no disease gene along chr $[0, l]$

H_1 : disease gene at $\tau \in [0, l]$

Define a *score function*

$$S : \{0, 1\}^m \rightarrow \mathbb{R}$$

so that $S(\mathbf{v})$ is large if \mathbf{v} is such that affected individuals share the same founder alleles. For *complete marker data*, define NPL process (Penrose, 1935, Whittemore and Halpern, 1994, Kruglyak et al, 1996)

$$Z(x) = S(\mathbf{v}(x)), \quad 0 \leq x \leq l,$$

where S is standardized so that

$$\begin{aligned} E_{H_0}(Z(x)) &= 2^{-m} \sum_{\mathbf{v}} S(\mathbf{v}) = 0 \\ \text{Var}_{H_0}(Z(x)) &= 2^{-m} \sum_{\mathbf{v}} S^2(\mathbf{v}) = 1. \end{aligned}$$

S is typically chosen so that

$$E_{H_1}(Z(x)) > 0 \text{ at all } x \text{ with max at } \tau,$$

N pedigrees

With N pedigrees, define

$$Z(x) = \sum_{k=1}^N \gamma_k S_k(\mathbf{v}_k(x)), \quad 0 \leq x \leq l,$$

with S_k and \mathbf{v}_k score fctn and inh vector for k^{th} pedigree and the weights $\gamma_k \geq 0$ satisfy

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

By standardization of all S_k

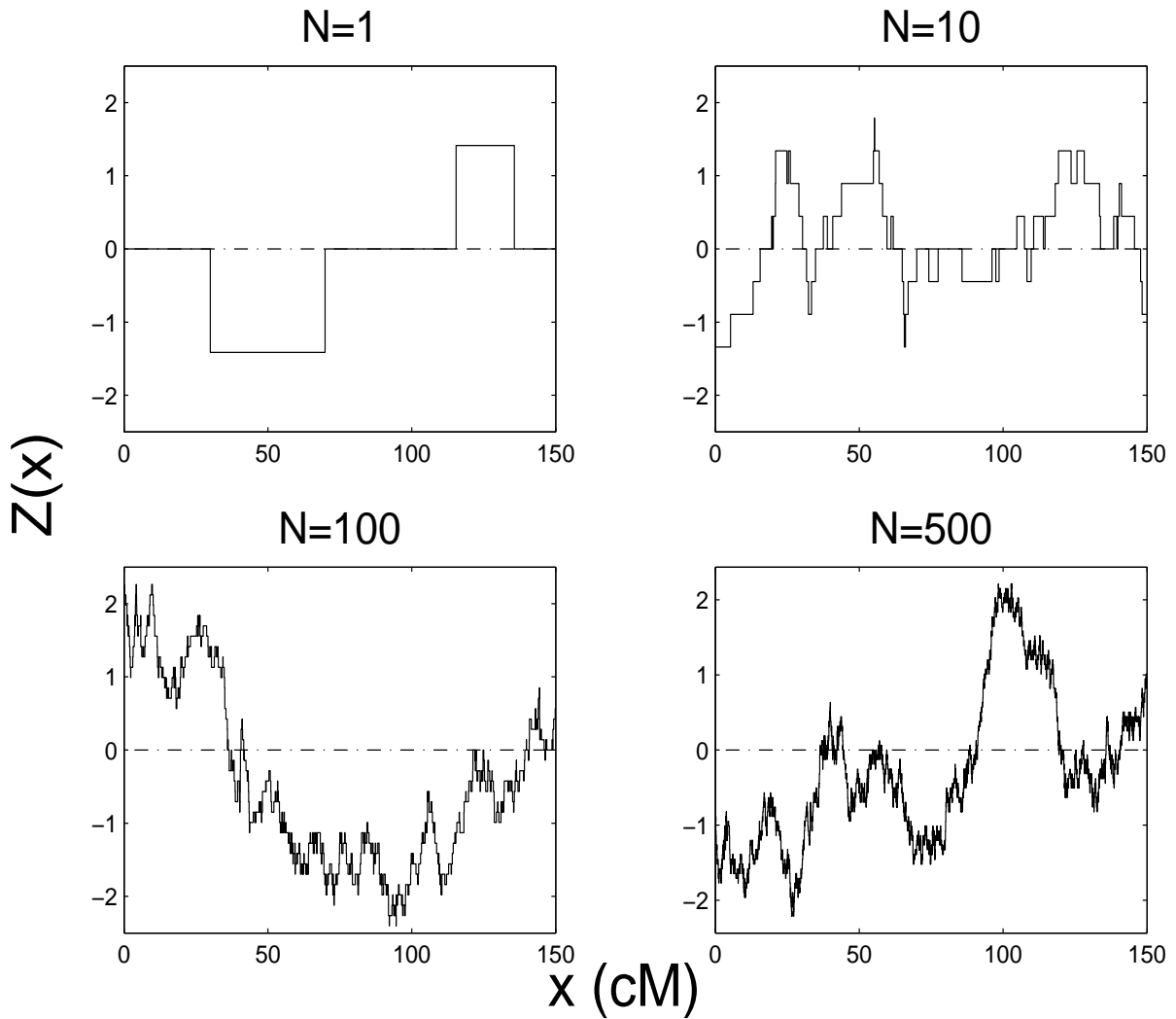
$$\begin{aligned} E_{H_0}(Z(x)) &= \sum_k \gamma_k \cdot 0 = 0 \\ \text{Var}_{H_1}(Z(x)) &= \sum_k \gamma_k^2 \cdot 1 = 1. \end{aligned}$$

$\{S_k\}$ are chosen so that

$$E_{H_1}(Z(x)) > 0 \text{ at all } x,$$

with maximum at $x = \tau$.

Plots of Z under H_0

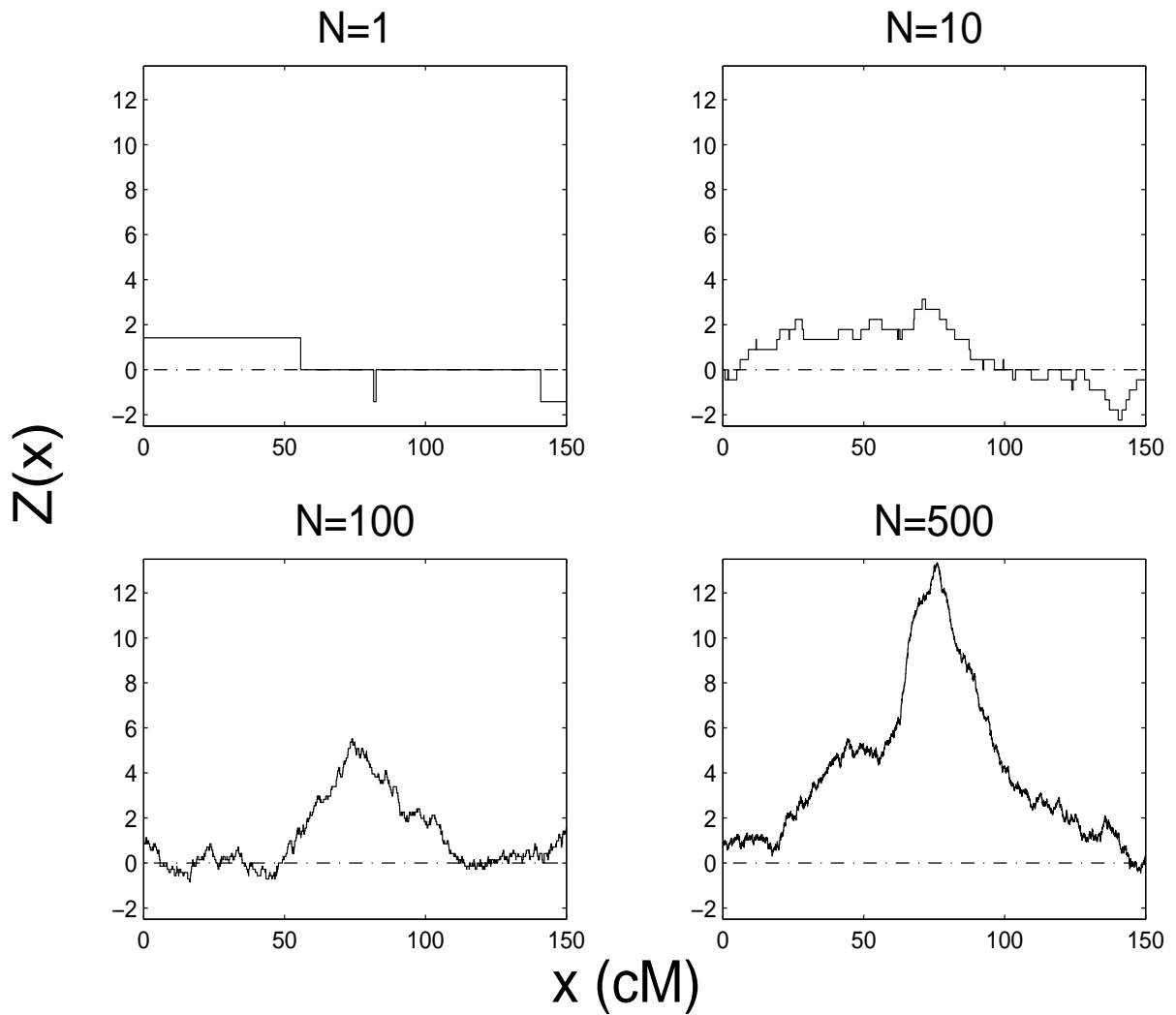


N identical pedigrees, $\gamma_k \equiv 1/\sqrt{N}$

Pedigree 1 (affected sib pair)

$$S(\mathbf{v}) = \sqrt{2}(1_{\{v_1=v_2\}} + 1_{\{v_3=v_4\}} - 1)$$

Plots of Z under H_1



$\tau = 75$ cM, Disease allele frequency 0.1
Penetrance parameters (0.05, 0.05, 1).

Test and Significance Level

Given threshold T , define test

$$Z_{\max} = \max_{0 \leq x \leq l} Z(x) \geq T \Rightarrow \text{reject } H_0.$$

We wish to compute the significance level (p -value)

$$\alpha = \alpha(T) = P_{H_0}(Z_{\max} \geq T).$$

With P denoting the distr of $Z = \{Z(x); 0 \leq x \leq l\}$ under H_0 , we rewrite this as

$$\alpha = \int f(z) dP(z),$$

where $f(z) = 1_{\{z_{\max} \geq T\}}$.

Methods of calculating α :

- 1. Analytical approximation:** Approximate Z by Gaussian Process, and use extreme value theory for 'continuous time' Gaussian processes. Lander and Bolstein (1989), Feingold et al (1993), Lander and Kruglyak (1995). Take non-Gaussianity into account by adjusting for skewness (Tang and Siegmund, 2001) or transform marginal to Gaussian (Änquist and Hössjer, 2004b).
+: Fast.
–: Difficult to generalize to incomplete marker data.
- 2. Direct Monte Carlo:** Generate i.i.d. replicates of Z under P (Boehnke, 1986, Ploughman and Boehnke, 1989).
+: Unbiased and consistent in limit in of many simulations.
–: Slow, especially for small α (10^{-6} or smaller).
- 3. Importance sampling:** Generate i.i.d. replicates of Z under \tilde{P} (Malley et al, 2002, Änquist and Hössjer, 2004a).
+: Unbiased and consistent in limit of many simulations and relatively fast, even for incomplete marker data and small α .
–: Can be slow for very large pedigrees..

Direct Monte Carlo:

We wish to compute

$$\alpha = E(f(Z)) = \int f(z)dP(z)$$

Direct Monte Carlo estimate:

$$\hat{\alpha} = \frac{1}{J} \sum_{j=1}^J f(Z_j),$$

where $Z_j \sim P$ are i.i.d. Hence $E(\hat{\alpha}) = \mu$ and $\text{Var}(\hat{\alpha}) = \sigma^2/J$, where

$$\sigma^2 = \int (f(z) - \alpha)^2 dP(z).$$

Importance Sampling

Rewrite

$$\alpha = \tilde{E}(f(Z)/L(Z)) = \int f(z)/L(z)d\tilde{P}(z),$$

where $L(z) = d\tilde{P}(z)/dP(z)$, provided $f(z)dP(z) > 0 \Rightarrow d\tilde{P}(z) > 0$. Estimate α by one of

$$\begin{aligned}\tilde{\alpha} &= \frac{1}{J} \sum_{j=1}^J f(Z_j)/L(Z_j), \\ \bar{\alpha} &= \frac{\sum_{j=1}^J f(Z_j)/L(Z_j)}{\sum_{j=1}^J 1/L(Z_j)},\end{aligned}$$

where $Z_j \sim \tilde{P}$ are i.i.d. The first estimate satisfies $\tilde{E}(\tilde{\alpha}) = \alpha$ and $\widetilde{\text{Var}}(\tilde{\alpha}) = \tilde{\sigma}^2/J$, where

$$\tilde{\sigma}^2 = \int (f(z)/L(z) - \alpha)^2 d\tilde{P}(z).$$

Kahn (1950). *Nucleonics* **6**(5), 27-37.

Kahn and Marshall (1953). *J. Oper. Res. Amer.* **1**, 263-278.

Hammersley and Handscomb (1964). *Monte Carlo Methods*.

Why Importance Sampling?

1. **Variance Reduction.** Sometimes, with proper choice of \tilde{P} , $\tilde{\sigma}^2 \ll \sigma^2$.
2. **Feasibility.** Sampling from P might be difficult or impossible.
3. **Reusing Samples** If several integrals

$$\alpha(T) = \int f(z; T) dP(z), \quad T \in \mathbb{T},$$

are of interest, sometimes a single sample $\{Z_j\} \sim \tilde{P}$ can be used for computing all $\alpha(T)$.

4. **A Generalization:** If the functions $f(\cdot; T)$ have almost disjoint support, one may sample from several \tilde{P} , or from a mixture distribution. (Geyer (1994), Hesterberg (1995)).

'Optimal' \tilde{P} :

In terms of **variance reduction**,

$$\tilde{P}(z) = f(z)dP(z)/\alpha \Rightarrow L(z) = f(z)/\alpha$$

is optimal, since $\tilde{\sigma}^2 = 0$. However, computing $L(\cdot)$ requires knowledge of α , the quantity we wish to estimate!!

Hence, the **computational gain** is a compromise between variance reduction and ease of simulating from \tilde{P} and computing $L(\cdot)$.

Construction of P

Under P (that is, under H_0), we generate Z as

1. **Select** $X = 0$.
2. For $k = 1, \dots, N$, **generate** $\mathbf{v}_k(X)$ from $P(\mathbf{v}_k(X) = \mathbf{v}) = 2^{-m_k}$, $m_k = \text{nr. of meioses of } k^{\text{th}} \text{ pedigree}$.
3. Let $\mathbf{v}_k(x) = (v_{k1}(x), \dots, v_{km_k}(x))$. For $k = 1, \dots, N$ and $j = 1, \dots, m_k$, **generate** independent Markov processes $\mathbf{v}_{kj}(\cdot) | \mathbf{v}_{kj}(X)$ with two states $\{0, 1\}$ and intensity matrix

$$\begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix}.$$

4. **Compute** $Z = \{Z(x), 0 \leq x \leq l\}$, where $Z(x) = \sum_{k=1}^N \gamma_k S_k(\mathbf{v}_k(x))$.

Alternatively, we might use $X \in U(0, l)$ in Step 1 and for each k in Step 3 generate two Markov processes *independently* to the left and right of X .

Generating $Z \sim \tilde{P}$

The construction is based on exponential tilting (Clark, 1966) and requires a tuning parameter $\delta \geq 0$:

1. **Generate** 'artificial disease locus' $X \in U(0, l)$,
2. For $k = 1, \dots, N$, **generate** $\mathbf{v}_k(X)$ from $\tilde{P}(\mathbf{v}_k(X) = \mathbf{v}) \propto \exp(\delta \gamma_k S_k(\mathbf{v}))$
3. **Generate** all $v_{kj}(\cdot) | v_{kj}(X)$ as under P , i.e. two independent Markov processes to the left and right of X .
4. **Compute** $Z = \{Z(x), 0 \leq x \leq l\}$, where $Z(x) = \sum_{k=1}^N \gamma_k S_k(\mathbf{v}_k(x))$.

OBSERVE: $\delta = 0$ yields simulation under P . $Z(X) \xrightarrow{\mathcal{L}} N(\delta, 1)$ under \tilde{P} as $N \rightarrow \infty$.

Likelihood Ratio

Write \tilde{P} as a mixture

$$d\tilde{P}(z) = \int_0^l d\tilde{P}_x(z) dx / l,$$

where \tilde{P}_x is \tilde{P} conditioned on $X = x$. Now

$$\begin{aligned} \frac{d\tilde{P}_x(z)}{dP(z)} &= \frac{\tilde{P}_x(z(x))}{P(z(x))} \\ &= \prod_{k=1}^N \frac{\tilde{P}_x(\mathbf{v}_k(x))}{P(\mathbf{v}_k(x))} \\ &\propto \prod_{k=1}^N \exp(\delta \gamma_k S_k(\mathbf{v}_k(x))) \\ &= \exp(\delta z(x)). \end{aligned}$$

Hence

$$\begin{aligned} L(z) &= \int_0^1 (d\tilde{P}_x(z) / dP(z)) dx / l \\ &= \int_0^l \exp(\delta z(x)) dx / (M(\delta)l), \end{aligned}$$

where $M(\delta) = E(\exp(\delta Z(x)))$ is a normalization constant (to ensure $\int d\tilde{P}(z) = 1$). See Frigessi and Vecellis (1985) and Naiman and Priebe (2001) for related algorithms (not using exponential tilting).

Several Thresholds

Based on i.i.d. $Z_j \sim \tilde{P}$ with parameter δ and $d\tilde{P}(z)/dP(z) = L(z; \delta)$, let

$$\tilde{\alpha}_\delta(T) = \frac{1}{J} \sum_{j=1}^J 1_{\{Z_{\max,j} \geq T\}} / L(Z_j; \delta).$$

The optimal choice of δ depends on T . If several T :s are of interest, we may use $M \geq 1$ values of $0 \leq \delta_1 \leq \dots \leq \delta_M$. Introduce

$$w = (w_1, \dots, w_M)$$

with $w_i \geq 0$ and $\sum_i w_i = 1$. Define

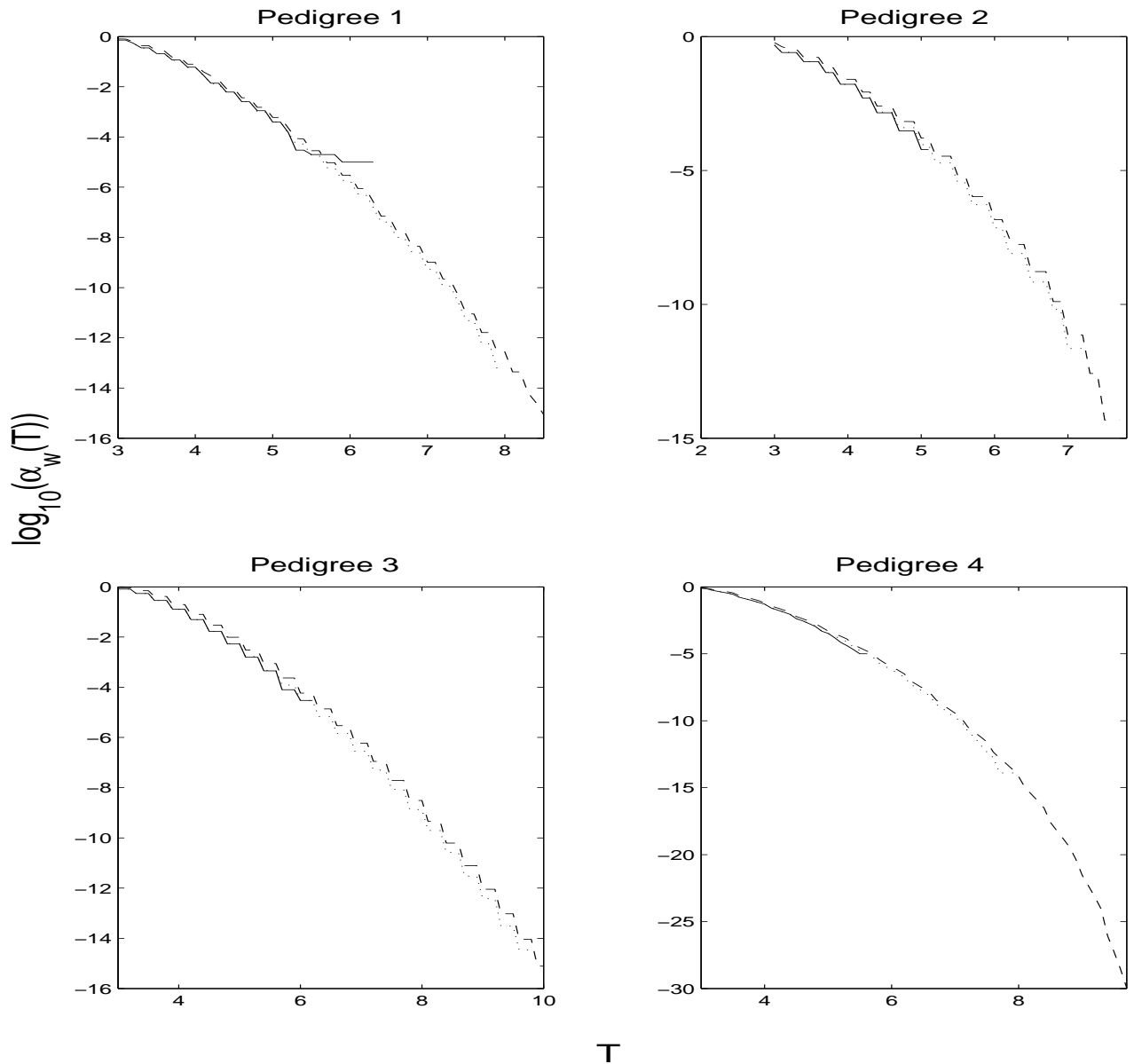
$$\tilde{\alpha}_w(T) = \sum_{i=1}^M w_i \tilde{\alpha}_{\delta_i}(T),$$

which satisfies $\tilde{E}(\tilde{\alpha}_w) = \alpha$ and $\tilde{\text{Var}}(\tilde{\alpha}_w) = \tilde{\sigma}^2 / J$, where

$$\tilde{\sigma}^2 = \sum_{i=1}^M w_i^2 \tilde{\sigma}_i^2,$$

with optimal $w_i \propto \tilde{\sigma}_i^{-1}$ estimated by plug-in estimate of each $\tilde{\sigma}_i$.

Computing $T \rightarrow \log_{10}(\alpha(T))$

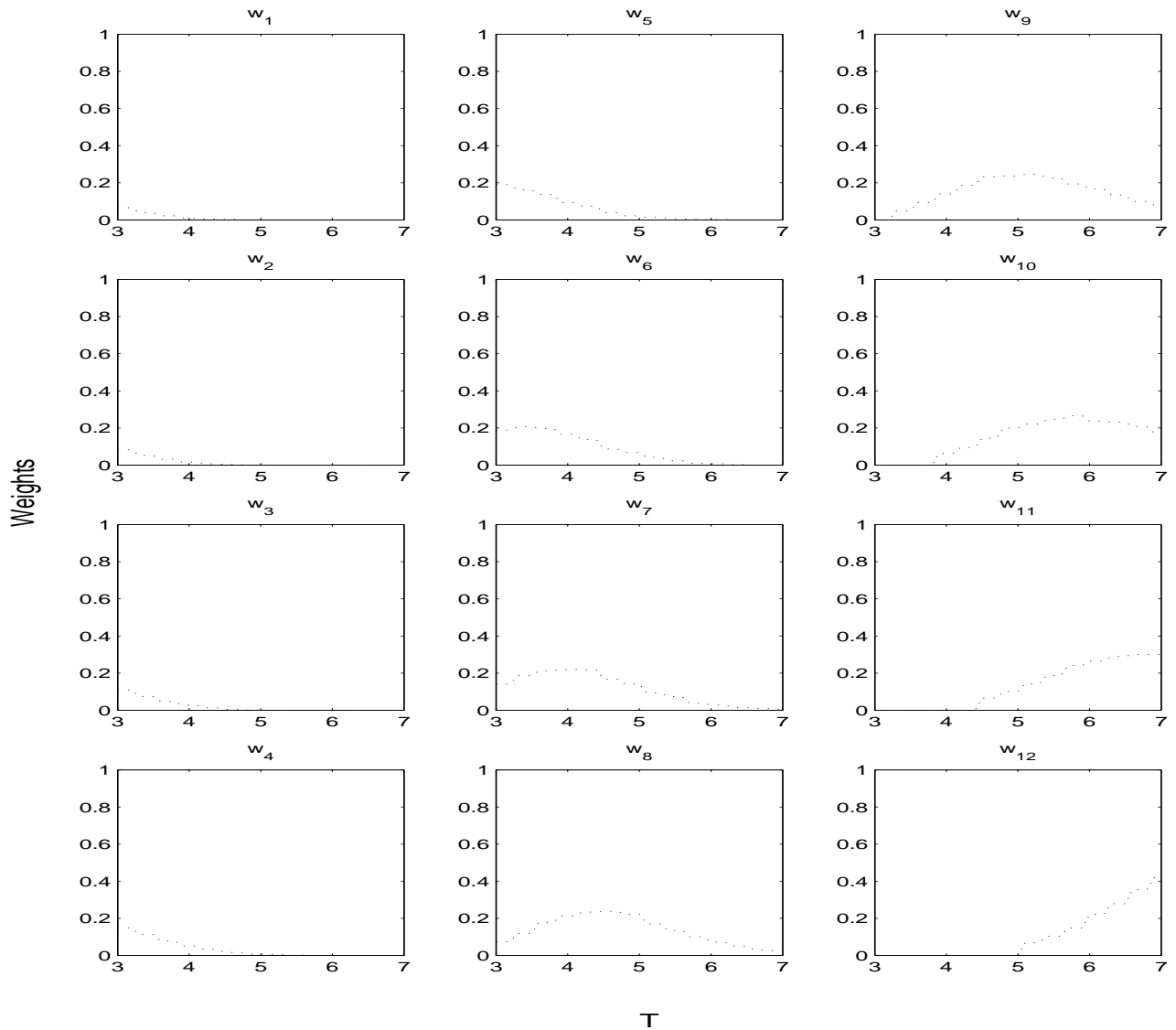


Genomewide scan: $\alpha(T) = 1 - \prod_{i=1}^{22} (1 - \alpha_i(T))$, $\sum_{i=1}^{22} l_i = 3575$ cM.

Solid: Direct MC, $J = 100.000$. Dashed: IS, $J = 3000$. Dotted: Analytical appr.

$N = 60$, $\gamma_k \equiv 1/\sqrt{N}$, $M = 12$, $\delta_0 = 0$, $\delta_{12} = 5.5$, $\{\delta_i\}$ equidistant, $CR \approx 3$

Displaying IS weights $w = w(T)$



$M = 12, w = (w_1, \dots, w_{12}), 3 \leq T \leq 7$

Pedigree 3, $J = 10.000, N = 60, \gamma_k \equiv 1/\sqrt{N}$.

Computational Gain

Is IS faster than Direct MC?

If the same number J of iterates is used for computing $\tilde{\alpha}_{\delta_i}$, $i = 1, \dots, M$, define the *Cost Adjusted Relative Efficiency*

$$\text{RE} = \frac{\sigma^2}{\tilde{\sigma}^2 \cdot \sum_{i=1}^M \text{CR}_i},$$

where

$\text{CR}_i = \text{Cost Ratio}$

is the time ratio to generate one $f(Z_j)/L(Z_j)$, $Z_j \sim \tilde{P}$ with $\delta = \delta_i$ in relation to one $f(Z_j)$, $Z_j \sim P$. In our case $\text{CR}_1 = 1$ and $\text{CR}_2 = \dots \text{CR}_M = \text{CR}$, assuming $0 = \delta_1 < \dots < \delta_M$.

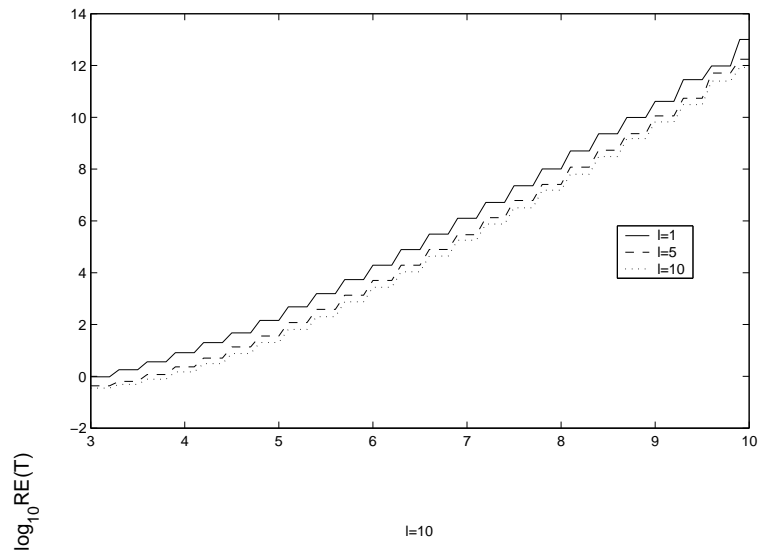
RE = relative time required for $\hat{\alpha}$ to achieve the same accuracy as $\tilde{\alpha}$.

Malley, Naiman and Bailey-Wilson. (2002). *Hum. Her.*, **54**, 174-185.

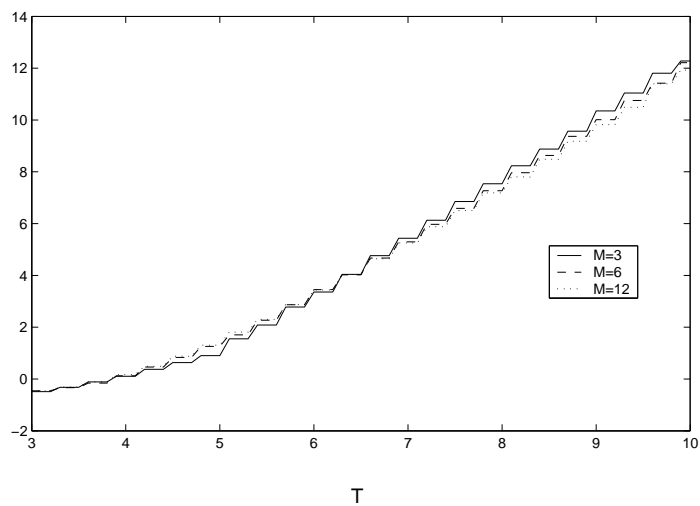
Änquist and Hössjer. (2004a). *Stat. Appl. Gen. Mol. Biol.*, to appear.

$T \rightarrow \log_{10}(\text{RE}(T))$ for Pedigree 3

M=12



l=10



$$N = 60$$

$$\gamma_k \equiv 1/\sqrt{N}$$

$\{\delta_i\}_{i=1}^M$ equispaced with $\delta_0 = 0, \delta_M = 5.5$

$S = S_{\text{all}}$ (Whittemore and Halpern, 1994),

Incomplete Marker Data

Let

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

be the NPL process for N families. When marker data is incomplete, we no longer observe family scores $Z_k(x) = S_k(\mathbf{v}_k(x))$, but

$$\begin{aligned} Z_k(x) &= E(S_k(\mathbf{v}_k(x)) | \text{MD}_k) \\ &= \sum_{\mathbf{v}} S_k(\mathbf{v}) P(\mathbf{v}_k(x) = \mathbf{v} | \text{MD}_k), \end{aligned} \quad (1)$$

where

$\text{MD}_k =$ marker data for k^{th} family,

see Kruglyak et al (1996). The probabilities in (1) are computed by means of a Hidden Markov forward-backward algorithm (Lander and Green, 1987).

Good news: It is possible to generalize the IS-estimator of $\alpha(T)$ to incomplete marker data.

Generate Marker Data under \tilde{P}

1. Given $\delta \geq 0$ and for $k = 1, \dots, N$, **generate** $\mathbf{v}_k(\cdot)$ under \tilde{P} as before (Steps 1-3 of alg. for complete marker data).
2. For $k = 1, \dots, N$ **generate** founder genotypes $\text{MD}_{k,\text{found}}$ for the k^{th} pedigree at all marker positions, according to marker allele frequencies.
3. For $k = 1, \dots, N$, **compute** MD_k as function of $\mathbf{v}_k(\cdot)$ and $\text{MD}_{k,\text{found}}$
4. **Compute** $Z = \{Z(x); 0 \leq x \leq l\}$ according to $Z(x) = \sum_{k=1}^N \gamma_k E(S_k(\mathbf{v}_k(x)) | \text{MD}_k)$.

IS Estimator, Incomplete Marker Data

The likelihood ratio is no longer a simple function of $Z = \{Z(x); 0 \leq x \leq l\}$, but of marker data

$$\mathbf{MD} = (\text{MD}_1, \dots, \text{MD}_k).$$

It can be shown (Ängquist and Hössjer, 2004b), that

$$\begin{aligned} L(\mathbf{MD}) &= \frac{d\tilde{P}(\mathbf{MD})}{dP(\mathbf{MD})} \\ &= \int_0^l \prod_{k=1}^N \sum_{\mathbf{v}} \exp(\delta \gamma_k S_k(\mathbf{v})) \\ &\quad \cdot P(\mathbf{v}_k(x) = \mathbf{v} | \text{MD}_k) dx / (lM(\delta)), \end{aligned}$$

which generalizes the complete marker data expression for $L(z)$. Based on i.i.d. $\mathbf{MD}_j \sim \tilde{P}$, the IS estimator (for one δ) is

$$\tilde{\alpha}(T) = \frac{1}{J} \sum_{j=1}^J 1_{\{Z_{\max,j} \geq T\}} / L(\mathbf{MD}_j).$$

Estimates for several δ can be weighted as for complete marker data.

Applications and Methodology of IS

- **Bootstrap** Resample non-uniformly from observations. Johns (1988), Davison (1988), Hinkley and Shi (1989).
- **Missing data** $Y =$ obs. data, $Z =$ missing data. i.i.d. replicates $Z_j \sim \tilde{P}$. Then

$$\frac{1}{J} \sum_{j=1}^J \frac{dP(Y, Z_j|\theta)}{d\tilde{P}(Z_j)}$$

is an unbiased and consistent estimate of $L(\theta) = P(Y|\theta)$.

- **Likelihood Ratios** If $d\tilde{P}(Z)$ is only known up to a normalizing constant, $dQ(z) \propto d\tilde{P}(Z)$, we can still produce a consistent estimate

$$\frac{\sum_j dP(Y, Z_j|\theta_2)/dQ(Z_j)}{\sum_j dP(Y, Z_j|\theta_1)/dQ(Z_j)}$$

of likelihood ratio $L(\theta_2)/L(\theta_1)$. See Ott (1979) and Geyer and Thompson (1994).

- **MCMC** One MCMC sample might be reused for several parameter values by importance sampling reweighting.

- **Risk Theory** To estimate ruin probability under P , sample from exponentially tilted random walk \tilde{P} , with ruin probability 1. See e.g. Ross (2000).
- **Population Genetics** Stephens and Donnelly (2002).
- **Compining several IS-samples.** Hesterberg (1995), Kong et al (2003).
- **Normalization methods of IS.** Hesterberg (1995).