Importance Sampling in Statistical Genetics

Efficient computation of $p$-values

Daniel Q. Naiman
daniel.naiman@jhu.edu

Department of Mathematical Sciences
Johns Hopkins University & Warwick University
Overview

* Introduction to genetics / affected sib pair test
* The multiple testing problem
* Simulation of stationary gaussian processes via DFT’s
* Importance sampling in general
* An importance sampling scheme
Essential Reading

Versellis & Frigessi (Union counting problem) - George Fishman’s Monte Carlo book


Carey Priebe, JHU Math Sciences (NP) Scan statistics paper in JCGS (2001)

James Malley, NIH & Joan Bailey-Wilson, CIDR - Manuscript in progress

Some Basic Genetics

* most people’s DNA comes in the form of 23 chromosome pairs

    maternal chromosome #1      paternal chromosome #1
    :                            :  
    maternal chromosome #22     paternal chromosome #22
    maternal chromosome #23     paternal chromosome #23

* chromosomes 1-22 are called autosomes (discussion below applies to these)

* 23rd (sex) chromosome is different and requires a different treatment

* chromosome pairs have similar structure

    linear & finite in size

    can talk about a locus (position) on each
Recombination

For each chromosome, each parent contributes to gamete (egg/sperm) a combination of their chromosome pairs.
Random Variables and Probabilistic Model Assumptions

* Consider 2 siblings, denoted by $S$ and $T$.
* Consider two loci $x$ and $y$ on the same chromosome.
* Introduce a random variable $X^{S,F}$ to represent the maternal grandparent responsible for the piece of DNA that the mother’s gamete contains.

\[
X^{S,F} = \begin{cases} 
F & \text{DNA from mother at locus } x \text{ from maternal grandmother} \\
M & \text{DNA from mother at locus } x \text{ maternal grandmother}
\end{cases}
\]

Similarly
* define random variables $X^{S,M}$, $X^{T,M}$, and $X^{T,F}$ for each parent, sib, at locus $x$.
* define random variables $Y^{S,F}$, $Y^{S,M}$, $Y^{T,F}$, and $Y^{T,F}$, for each parent, sib, at locus $y$.

Assumptions:
* for any locus, there is an equal chance of maternal DNA coming from either maternal grandparent.
* for any locus, there is an equal chance of paternal DNA coming from either paternal grandparent.
* maternal DNA composition is independent of paternal DNA composition.
* process of combining chromosome pairs is independent from chromosome to chromosome.
* process of composition between siblings (sibs) is independent.
Consequences of the assumptions

- $P[X^{S,F} = M] = \frac{1}{2}$ and $P[X^{S,F} = F] = \frac{1}{2}$
- $P[Y^{S,F} = M] = \frac{1}{2}$ and $P[Y^{S,F} = F] = \frac{1}{2}$
- $X^{S,M}, X^{S,F}, X^{T,M},$ and $X^{T,F}$ are all independent
- $Y^{S,M}, Y^{S,F}, Y^{T,M},$ and $Y^{T,F}$ are all independent

The only source of dependence: same parent, same sib, possibly different locus:

- $X^{S,F}$ and $Y^{S,F}$ are dependent
- $X^{S,M}$ and $Y^{S,M}$ are dependent
- $X^{T,F}$ and $Y^{T,F}$ are dependent
- $X^{T,M}$ and $Y^{T,M}$ are dependent
Dependence, Recombination Distance and Allele Sharing

Recombination distance: When pieces of DNA break and recombine in production of gametes, we can define for a pair of loci on the same chromosome the probability $\theta$ that the DNA at locus $x$ and $y$ come from the different grandparents (not identical by descent)

$$\theta = \theta(x, y) = P[X^{S,F} \neq Y^{S,F}] = P[X^{S,M} \neq Y^{S,M}].$$

Allele sharing: For the sib pair $S$ and $T$, we define $N^x$ to be the number of alleles shared IBD at locus $x$, that is, the number of pieces of DNA in a neighborhood of $x$ originating from the same grandparent.

$$N^x = N^{x,F} + N^{x,M}$$

where

$$N^{x,F} = I(X^{S,F} = X^{T,F})$$

the indicator that the maternal DNA at $x$ comes from the same grandparent, and

$$N^{x,M} = I(X^{S,M} = X^{T,M})$$

indicator that the paternal DNA at $x$ comes from the same grandparent. In a similar manner, define $N^{y,F}$ and $N^{y,M}$. 
Moments of $N^x$

Key points:

* $N^x,F$ is a Bernoulli random variable
* $N^x,M$ is a Bernoulli random variable
* $N^x,M$ and $N^x,F$ are independent

\[
E[N^x] = 1
\]

\[
\text{Var}(N^x) = 1/2
\]
Affected Sib Pair (ASP) Test

* Assume an inherited disease
* Assume DNA at particular disease locus causes the disease
* Assume a collection of affected sib pairs \((S_1, T_1), (S_2, T_2), \ldots, (S_P, T_P)\) (pairs unrelated)
* Assume we can determine IBD allele sharing at a marker locus \(x\)
* Let \(N^x_j\) = sharing of alleles IBD at locus \(x\) for sib pair \((S_j, T_j)\).
* Define \(N^x = \sum_{j=1}^{P} N^x_j\) = total sharing of alleles IBD by sib pairs at locus \(x\).

Idea of the ASP test: If marker locus \(x\) is close to the disease locus, we expect greater IBD allele sharing than would be obtained by chance alone.
Test based on a particular marker $x$

**Testing problem:** $H$: disease and marker $x$ unlinked vs. $A$: disease and marker $x$ linked by rejecting $H$ if $N^x$ is sufficiently large.

**Moments of $N^x$:**

\[
E[N^x] = P
\]

\[
\text{Var}(N^x) = P/2
\]

**z-test:** Reject $H$ if

\[
Z_x = \frac{(N^x - P)}{\sqrt{P/2}} \geq z_\alpha,
\]

where $z_\alpha$ denotes the upper $\alpha$ critical point for the $N(0, 1)$ distribution.

**False rejection probability:** If $H$ holds, then the probability of rejection is approximately $\alpha$, by the central limit theorem.
Multiple markers & the problem of multiple tests

* Suppose we have several markers $x_1, \ldots, x_m$ and we perform this same test for each marker

* Now we have a whole family of hypothesis tests to perform

$$H_i : \text{disease and marker } x_i \text{ unlinked } \text{ vs. } A_i : \text{disease and marker } x_i \text{ linked}$$

* A priori we may have no information as to the relative plausibility of the several hypotheses, so the inference involves searching through the family of hypotheses for an index $i$ such that a particular $A_i$ is plausible, and testing $H_i$ vs. $A_i$

* Can reject $H_i$ (conclude linkage at marker $x_i$) if $N^{x_i}$ is sufficiently large for some $i$, say if

$$Z_{x_i} \geq C$$

* We would like to be able to control the probability of at least one false rejection, i.e. we want

$$P[Z_{x_i} \geq C, \text{ for some } i] \leq \alpha,$$

where $\alpha \in (0, 1)$ is prescribed.
Bonferroni method/correction

Typical approach to controlling the overall error rate for \( m \) tests is to take \( C = z_\alpha / m \) so that

\[
P[Z_{x_i} \geq C, \text{ for some } i] = P[\bigcup_{i=1}^{m} E_i] \leq \sum_{i=1}^{m} P[E_i] = m \alpha / m = \alpha
\]

* Method is exact if the events \( E_i \) are disjoint, but otherwise conservative

* Problem: approximate the error probability, i.e. determine distribution of \( \max_{i=1,\ldots,m} Z_{x_i} \)

* First need the (approximate) joint distribution of the \( Z_{x_i} \)

* Since \( Z_{x_i} \) is a normalized average of iid random variables, the multivariate central limit theorem tells us that \( (Z_{x_1}, \ldots, Z_{x_m}) \) is approximately \( \mathcal{N}(0, \Sigma) \), where the \( \Sigma = (\sigma_{ij}) \) with

\[
\sigma_{ij} = \text{Cov}(Z_{x_i}, Z_{x_j}) = \text{Cov}(N_{x_i}^{x_i}, N_{x_j}^{x_j})
\]

the covariances between allele sharing for single sib pairs
Covariance between allele sharing at loci \( x \) & \( y \) (single sib pair)

Observe that the random variables \( N^{x,F}, N^{x,M}, N^{y,F}, \) and \( N^{y,M} \) are identically distributed with

\[
N^{\cdot,\cdot} = \begin{cases} 
0 & \text{w.p. } \frac{1}{2} \\
1 & \text{w.p. } \frac{1}{2} 
\end{cases}
\]

Furthermore

\[
\ast \quad N^{x,F} \perp N^{x,M}, \quad N^{y,F} \perp N^{y,M}, \quad N^{x,F} \perp N^{y,M}, \quad \text{and} \quad N^{x,M} \perp N^{y,F}
\]

Now

\[
\text{Cov}(N^{x,F}, N^{y}) = \text{Cov}((N^{x,F} + N^{x,M}), (N^{y,F} + N^{y,M}))
\]

\[
= \text{Cov}(N^{x,F}, N^{y,F}) + \text{Cov}(N^{x,F}, N^{y,M}) + \text{Cov}(N^{x,M}, N^{y,F}) + \text{Cov}(N^{x,M} N^{y,M})
\]

\[
= 2 \text{Cov}(N^{x,F}, N^{y,F})
\]

\[
= 2 \left( P[N^{x,F} = 1, N^{y,F} = 1] - P[N^{x,F} = 1]P[N^{y,F} = 1] \right)
\]

\[
= 2 \left( P[N^{x,F} = 1, N^{y,F} = 1] - \frac{1}{4} \right)
\]
Covariance calculation continued

\[
P \left[ N^x, F = 1, N^y, F = 1 \right] = P \left[ X^{S, F} = X^{T, F}, Y^{S, F} = Y^{T, F} \right]
\]

\[
\]

\[
= P \left[ (X^{S, F}, Y^{S, F}) = (F, F), (X^{T, F}, Y^{T, F}) = (F, F) \right]
\]

\[
+ P \left[ (X^{S, F}, Y^{S, F}) = (F, M), (X^{T, F}, Y^{T, F}) = (F, N) \right]
\]

\[
+ P \left[ (X^{S, F}, Y^{S, F}) = (M, F), (X^{T, F}, Y^{T, F}) = (M, F) \right]
\]

\[
+ P \left[ (X^{S, F}, Y^{S, F}) = (M, M), (X^{T, F}, Y^{T, F}) = (M, M) \right]
\]
Covariance calculation continued

\[ P \left[ (X^{S,F}, Y^{S,F}) = (F, F) \right] P \left[ (X^{T,F}, Y^{T,F}) = (F, F) \right] \]

\[ + P \left[ (X^{S,F}, Y^{S,F}) = (F, M) \right] P \left[ (X^{T,F}, Y^{T,F}) = (F, M) \right] \]

\[ + P \left[ (X^{S,F}, Y^{S,F}) = (M, F) \right] P \left[ (X^{T,F}, Y^{T,F}) = (M, F) \right] \]

\[ + P \left[ (X^{S,F}, Y^{S,F}) = (M, M) \right] P \left[ (X^{T,F}, Y^{T,F}) = (M, M) \right] \]

\[ = \frac{1}{2} (1 - \theta) \times \frac{1}{2} (1 - \theta) + \frac{1}{2} \theta \times \frac{1}{2} \theta + \frac{1}{2} \theta \times \frac{1}{2} \theta + \frac{1}{2} (1 - \theta) \times \frac{1}{2} (1 - \theta) \]

\[ = \frac{1}{2} (1 - \theta)^2 + \frac{1}{2} \theta^2. \]

Substituting above, we see that

\[ \text{Cov}(N^X, N^Y) = (1 - \theta)^2 + \theta^2 - \frac{1}{2} \]
Haldane’s map function

* Gamete production: chromosomes replicate to give 4 strands, 2 maternal & 2 paternal. They break at random positions (called chiasmata) and recombine.
* The recombination distance (not really a distance) between 2 loci tells us the probability that the number of chiasmata) between 2 loci is odd.
* Define $d$ expected number of chiasmata between the two loci. Then under suitable model assumptions the recombination distance satisfies $\theta = \frac{1}{2} (1 - \exp\{-d/2\})$, and in terms of $d$ we see that the above covariance becomes $\text{Cov}(N^x, N^y) = \frac{1}{2} \exp\{-d\}$. Since $\text{Var}(N^x) = \text{Var}(N^X) = \frac{1}{2}$ we see that

$$\rho(N^X, N^Y) = \exp\{-d\}$$
Restatement of the problem

* Fix a collection of points $x_1, \ldots, x_d$ in the real line whose pairwise distance is $d(x_i, x_j)$.

* Let $(Z_{x_1}, \ldots, Z_{x_m})$ be a multivariate normal random vector with mean 0 and covariance matrix given by

$$\text{Cov}(Z_{x_i}, Z_{x_j}) = \exp\{-d_{x_i, x_j}\}.$$  

* Find the distribution of $\max_{i=1,\ldots,m} Z_{x_i}$. i.e.

$$p = P[\max_{i=1,\ldots,m} Z_{x_i} \geq C]$$  

for given constant $C$. 
Naive Monte Carlo Simulation Approach

* Generate a sample $Z^{(j)} = (Z_{x_1}^{(j)}, \ldots, Z_{x_m}^{(j)})$ with having a $N(0, \Sigma)$ distribution, for $j = 1, \ldots, n$.

* Estimate using

$$
\hat{p} = \frac{1}{n} \sum_{j=1}^{n} I \left( \max_{i=1,\ldots,m} Z_{x_i}^{(j)} \geq C \right)
$$

Two key issues to address

* How to efficiently generate a sample?

* What happens if $p$ is very small?
Generating multivariate Gaussian random variables

Standard method for generating \( Z \sim N_m(0, \Sigma) \):

* Find Cholesky decomposition of \( \Sigma \):
  \[
  CC^t = \Sigma
  \]

* Generate \( X \sim N(0, I_m) \) i.e. \( X \) has independent standard normal components.

* Take \( Z = CX \).

**Issue:** \( m \) might be huge and each \( Z \) requires \( O(m^2) \) effort to perform the matrix multiplication

* Can exploit stationarity and speed of fast fourier transforms to do much better
Relative inefficiency of naive Monte Carlo for small $p$

Naive Monte Carlo estimate of $p = P[Z \in A]$: 
$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} I(Z^{(i)} \in A)$$

$$n\hat{p} \sim \text{Binomial}(n, p)$$

$$E[\hat{p}] = p, \text{ i.e. } \hat{p} \text{ is unbiased}$$

$$\text{Var}(\hat{p}) = p(1 - p)/n$$

Relative variability:

$$\frac{\text{SE}(\hat{p})}{p} = \sqrt{\frac{1 - p}{np}} \approx \frac{1}{\sqrt{np}} \to \infty \text{ as } p \to 0.$$ 

Importance sampling avoids this problem
Exploiting stationarity

Essential ideas:

* Embed the sequence to be generated $Z_{x_1}, \ldots, Z_{x_m}$ in a sequence $Z_x, x \in G$ where $G$ is an equispaced grid.

* Covariance matrix now has Toeplitz structure

\[
\Sigma = \begin{pmatrix}
1 & r_1 & r_2 & \cdots & r_{m-2} & r_{m-1} \\
r_1 & 1 & r_1 & \cdots & r_{m-3} & r_{m-2} \\
& & \ddots & \ddots & \ddots & \ddots \\
& & & 1 & r_1 & r_{m-1} \\
& & & & r_{m-2} & r_{m-3} \\
& & & & & r_{m-1}
\end{pmatrix}
\]

* Embed this matrix in a larger one $\Gamma$ so that $\Sigma$ is the upper left hand corner, where $\Gamma$ is circulant.
Minimal embedding of Toeplitz in circulant

\[
\begin{pmatrix}
1 & r_1 & r_2 & \cdots & r_{m-2} & r_{m-1} & r_{m-2} & \cdots & r_2 & r_1 \\
 r_1 & 1 & r_1 & \cdots & r_{m-3} & r_{m-2} & r_{m-1} & \cdots & r_3 & r_2 \\
 \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
 r_2 & r_3 & r_4 & \cdots & r_{m-2} & r_{m-3} & r_{m-4} & \cdots & 1 & r_1 \\
 r_1 & r_2 & r_3 & \cdots & r_{m-1} & r_{m-2} & r_{m-3} & \cdots & r_1 & 1
\end{pmatrix}
\]

* This might not be positive semidefinite in general, but Theorem due to Dietrich & Newsome says it is as long as the correlation function is convex & decreasing (which is true in the exponential case)

* If it is, a random vector with this covariance matrix is stationary with index set being the set of integers modulo $2m$
The discrete Fourier transform (DFT)

Given a finite sequence of \( N \) real numbers \( x = (x_0, \ldots, x_{N-1}) \) we define its forward DFT

\[
\hat{x}_k = \sum_{j=0}^{N-1} \exp\{2\pi i j k / N\}
\]

and its backward DFT by

\[
\tilde{x}_k = \frac{1}{N} \sum_{j=0}^{N-1} \exp\{-2\pi i j k / N\}
\]
Facts about the discrete Fourier transform (DFT)

* \( \tilde{x} = x \)

* DFT’s can be computed in \( O(N \log N) \) time (the FFT) (see FFTW software)

* Can compute a Cholesky decomposition of positive definite circulant matrix using DFT:
  
  forward transform the first row
  compute square root componentwise
  backward transform
  form the circulant matrix using the result

* Circulant matrix times vector multiplication \( \Gamma u \) simplifies
  
  forward transform \( u \)
  forward transform first row \( \gamma \) of \( \Gamma \)
  pointwise multiply \( \hat{\gamma} \) and \( \hat{u} \)
  back transform
Importance sampling in general

Problem: Approximate $\mu = \int g(x)f(x)dx$ where $f$ is a probability density function.

Naive procedure: Generate $X_1, \ldots, X_N$ iid with $X_i \sim f$ and take

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} g(X_i)$$

Properties: $E[\hat{\mu}] = \mu$ and $\text{Var}(\hat{\mu}) = \frac{1}{N} \text{Var}_f(X_i)$

Importance sampling: Write $\mu = \int \tilde{g}(x)\tilde{f}(x)dx$, where $\tilde{g} = gf/\tilde{f}$ and $\tilde{f}$ is a pdf with

$$\{\tilde{f} = 0\} \subseteq \{gf = 0\}$$

Importance sampling procedure: Generate $X_1, \ldots, X_N$ iid with $X_i \sim \tilde{f}$ and take

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} \tilde{g}(X_i)$$

Properties: $E[\tilde{\mu}] = \mu$ and $\text{Var}(\tilde{\mu}) = \frac{1}{N} \text{Var}_{\tilde{f}}(X_i)$

Special choice of $\tilde{f}$: If $g \geq 0$ and we take $\tilde{f} \propto gf$ then $\text{Var}(\tilde{\mu}) = 0$

However: Sampling from $\tilde{f}$ may need normalization constant. Knowing this means knowing $\mu$. 
Importance sampling for correction to Bonferroni

Observe a random vector $Z$ and want to determine $P[\bigcup_{i=1}^{m} \{ Z \in E_i \} ]$

Bonferroni bound: $B = \sum_{i=1}^{m} P[ Z \in E_i ]$

Derivation of a correction to Bonferroni:

\[
P \left[ \bigcup_{i=1}^{m} \{ Z \in E_i \} \right] = \int I_{\bigcup_{i=1}^{m} \{ Z \in E_i \}} dP
\]

\[
= \int \frac{I_{\bigcup_{j=1}^{N} \{ Z \in E_j \}}}{\sum_{j=1}^{N} I_{\{ Z \in E_j \}}} \sum_{i=1}^{m} I_{\{ Z \in E_i \}} dP
\]

\[
= B \sum_{i=1}^{m} q_i \int \frac{1}{g(Z)} \frac{I_{\{ Z \in E_i \}}}{P[Z \in E_i]} dP
\]

where

\[
g(Y) = \sum_{i=1}^{m} I_{\{ Z \in E_i \}}
\]

\[
q_i = P[Z \in E_i]/B.
\]
**Consequences**

We have

\[
P[\bigcup_{i=1}^{m} \{Z \in E_i\}] = B\rho
\]

where

\[
\rho = \sum_{i=1}^{m} q_i \int \frac{1}{g(Z) P[Z \in E_i]} dP.
\]

**Interpretation of correction factor \( \rho \):** expected value of random variable \( \hat{\rho} \) generated in a single iteration of the following Monte Carlo experiment

**Step 1.** Generate index \( J \in \{1, \cdots, m\} \) according to the \( q_i \).

**Step 2.** Generate \( \tilde{Z} \) from the conditional distribution \( Z \) given that \( Z \in E_J \).

**Step 3.** Count \( g = g(\tilde{Z}) \) the number of \( i \) s.t. \( \tilde{Z} \in E_i \).

Take \( \hat{\rho} = 1/g \).
How good is our approximation?

Estimate of $p$: $\hat{p} = B\hat{\rho}$, where $\hat{\rho}$ is based on some Monte Carlo sample size $N$

Upper bound for $\text{Var}(\hat{p})$: $\hat{\rho}$ is an average of iid random variables, $0 \leq 1/g \leq 1$, so its variance is bounded above by $1/4N$

$$\text{Var}(\hat{p}) \leq \frac{B^2}{4N}$$

Typical implementation issue: Can we sample from the conditional distribution easily?

Genetics application: Yes. Not only that, we can represent a conditional sample as filtered white noise, and use FFT’s to generate it FAST!!!

Follows from elementary matrix analysis arguments:

- assume Cholesky decomposition for covariance matrix
- write down the conditional distribution
  
  : 
  : 
  : 
  : 
  : 
  : 
  :
**Numerical results**

Markers equally spaced over genome

**Relative efficiency of importance sampling vs. naive sampling** ratio of computation time that achieves estimates with equal variance

**Exceedance threshold**: Bonferroni with $\alpha = .05$

<table>
<thead>
<tr>
<th>Number of markers</th>
<th>Bonferroni</th>
<th>$p$</th>
<th>Relative eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>.05</td>
<td>.029</td>
<td>79.5</td>
</tr>
<tr>
<td>64</td>
<td>.05</td>
<td>.021</td>
<td>72.7</td>
</tr>
<tr>
<td>128</td>
<td>.05</td>
<td>.014</td>
<td>59.3</td>
</tr>
<tr>
<td>256</td>
<td>.05</td>
<td>.0089</td>
<td>67.6</td>
</tr>
<tr>
<td>512</td>
<td>.05</td>
<td>.0055</td>
<td>59.1</td>
</tr>
<tr>
<td>1,024</td>
<td>.05</td>
<td>.0032</td>
<td>84.9</td>
</tr>
<tr>
<td>2,048</td>
<td>.05</td>
<td>.0019</td>
<td>95.7</td>
</tr>
<tr>
<td>4,096</td>
<td>.05</td>
<td>.0010</td>
<td>122.6</td>
</tr>
<tr>
<td>8,192</td>
<td>.05</td>
<td>.00058</td>
<td>209.5</td>
</tr>
<tr>
<td>16,384</td>
<td>.05</td>
<td>.00032</td>
<td>208.8</td>
</tr>
</tbody>
</table>
Numerical results

Markers equally spaced over genome

Relative efficiency of importance sampling vs. naive sampling ratio of computation time that achieves estimates with equal variance

Exceedance threshold: Bonferroni with $\alpha = .01$

<table>
<thead>
<tr>
<th>Number of markers</th>
<th>Bonferroni</th>
<th>$p$</th>
<th>Relative eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>.01</td>
<td>.0066</td>
<td>520</td>
</tr>
<tr>
<td>64</td>
<td>.01</td>
<td>.0049</td>
<td>431</td>
</tr>
<tr>
<td>128</td>
<td>.01</td>
<td>.0033</td>
<td>325</td>
</tr>
<tr>
<td>256</td>
<td>.01</td>
<td>.0021</td>
<td>392</td>
</tr>
<tr>
<td>512</td>
<td>.01</td>
<td>.0013</td>
<td>225</td>
</tr>
<tr>
<td>1,024</td>
<td>.01</td>
<td>.00076</td>
<td>375</td>
</tr>
<tr>
<td>2,048</td>
<td>.01</td>
<td>.00043</td>
<td>396</td>
</tr>
<tr>
<td>4,096</td>
<td>.01</td>
<td>.00024</td>
<td>625</td>
</tr>
<tr>
<td>8,192</td>
<td>.01</td>
<td>.00013</td>
<td>682</td>
</tr>
<tr>
<td>16,384</td>
<td>.01</td>
<td>.000072</td>
<td>709</td>
</tr>
</tbody>
</table>
Final comments

* Method handles unequally spaced markers easily

* Conditional sampling algorithm can be extended to lots of non-Gaussian cases (random variables constructed from Gaussians)

* Extreme value theory also gives answers:

  haven’t compared extensively
  don’t know how robust the theory is

* Lots of work yet to be done: applications to other problems in genetics, e.g. association tests, transmission disequilibrium tests, . . .

* Thanks for listening