

Asymptotic Estimation Theory of Multipoint Linkage Analysis Under Perfect Marker Information

Ola Hössjer
Lund University, Sweden

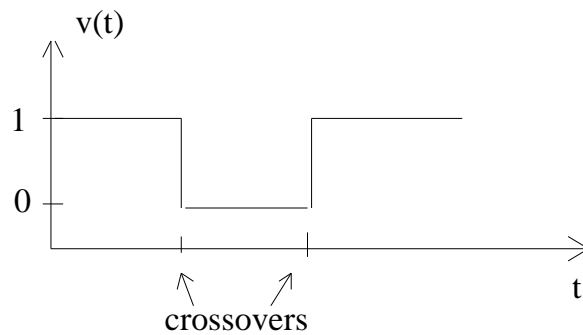
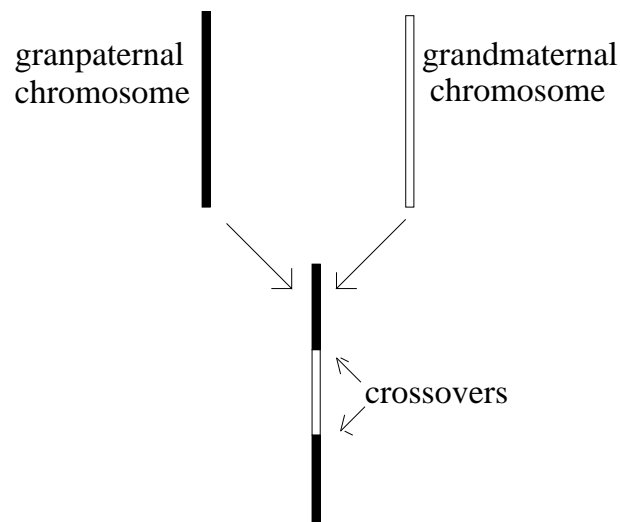
June 4, 2002

Linkage analysis

- **Objective:** Locate position τ of a gene causing (contributing to) a disease along a chromosome.
- **Phenotype data:** Affection status (or other variables) from N pedigrees, i.e. families susceptible to the disease.
- **Marker data:** Trace inheritance of a number of marker genes (at known positions) through the pedigrees.
- **Estimate τ :** Maximize a score function $t \rightarrow Z_N(t)$ along the chromosome. $Z_N(t)$ (e.g. log likelihood) measures 'agreement' between inheritance at locus (position) t (found from marker genes) and inheritance of disease phenotypes.

Crossovers

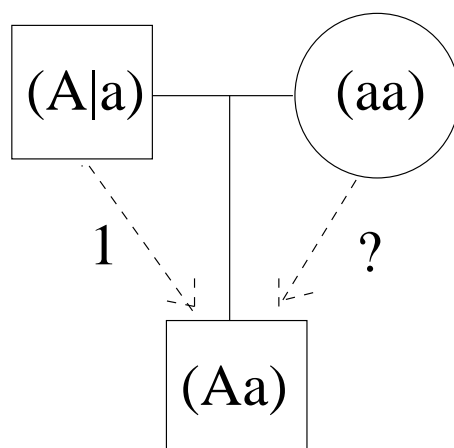
Formation of one gamete (from father or mother) can be described as follows:



The process is called a *meiosis*. On average 33 crossovers occur along 23 chromosomes.

Known inheritance

Let A and a denote the disease causing and normal allele respectively at the disease locus τ . Three possible genes; (AA) , (Aa) and (aa) can be formed. A gene $(A|a)$ with known phase has the first allele (A) transmitted from the father and the second one (a) from the mother.



In the above example, $v(\tau) = 1$ for the paternal meiosis and $v(\tau)$ is unknown for the maternal meiosis.

Inheritance process (one meiosis)

$\{v(t); 0 \leq t \leq l\}$, where $v(t) = 1 (= 0)$ if parental (maternal) allele is transmitted at locus t .

Haldane's map fctn: Crossovers occur acc. to Poisson process with intensity λ ($=0.01$ if map distance in cM)
 $\implies v(\cdot)$ is a 2-state Markov process with intensity matrix

$$\begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}$$

Recombination occurs between t and $t + h$ if $v(t) \neq v(t + h)$ (odd number of crossovers between t and $t + h$)

Recombination fraction between t and $t + h$ is

$$\begin{aligned} \theta_h &= P(\text{recombination between } t \text{ and } t + h) \\ &= \frac{1}{2} - \frac{1}{2} \exp(-2\lambda h). \end{aligned}$$

Single point analysis

Objective: Locate position of disease locus τ .

Suppose we have one marker gene at t_0 , which is fully informative, i.e. $v(t_0)$ can be observed.

Observed data (phenotype): $Y = v(\tau)$. The likelihood for one meiosis is

$$L(t) = P(Y|v(t_0), \tau = t) = \begin{cases} \theta_{|t-t_0|}, & Y \neq v(t_0), \\ 1 - \theta_{|t-t_0|}, & Y = v(t_0). \end{cases}$$

For a data set of N meioses, the likelihood becomes:

$$L(t) = \prod_{i=1}^N L_i(t) = \theta_{|t-t_0|}^X (1 - \theta_{|t-t_0|})^{N-X},$$

where L_i is the likelihood for the i :th meiosis and X is the number of observed recombinations between t_0 and τ in the N meioses. $X \in \text{Bin}(N, \theta_{|t-t_0|})$. ML estimate of τ is $\hat{\tau}_N =$ solution w.r.t. t to $\theta_{|t-t_0|} = X/N$ (not unique). One 'version' of solution \sqrt{N} -consistent and as. normal.

Two point analysis (one meiosis)

Suppose we have two markers at t_0 and t_1 , such that marker data $\text{MD} = (v(t_0), v(t_1))$ is observed. Then, if $t_0 < t < t_1$, the likelihood becomes

$$L(t) = P(Y|\text{MD}, \tau = t) = \begin{cases} P_{00}(t); & \text{MD} = (Y, Y) \\ P_{01}(t); & \text{MD} = (Y, 1 - Y) \\ P_{10}(t); & \text{MD} = (1 - Y, Y) \\ P_{11}(t); & \text{MD} = (1 - Y, 1 - Y) \end{cases}$$

where $Y = v(\tau)$,

$$P_{11}(t) = \frac{P(v(t_0) \neq Y \neq v(t_1))}{P(v(t_0) = v(t_1))} = \frac{\theta_{|t-t_0|} \theta_{|t-t_1|}}{\theta_{|t_1-t_0|}}$$

and so on.

Two point analysis, contd.

With N meioses, let X_{ij} denote the number of events corresponding to 'case P_{ij} '. Put $X = \{X_{ij}\}_{i,j=0}^1$ and $P(t) = \{P_{ij}(t)\}_{i,j=0}^1$. Then, the likelihood is

$$L(t) = P(t)^X = \prod_{i,j=0}^1 P_{ij}(t)^{X_{ij}},$$

and

$$\hat{\tau}_N = \arg \max_t L(t)$$

is \sqrt{N} -consistent and asymptotically normal, with asymptotic variance depending on (t_0, τ, t_1) .

Perfect marker information

With an infinitely dense set of markers, we assume that $MD = \{v(t); 0 \leq t \leq l\}$ is observed. Assume w.l.o.g. $Y = v(\tau) = 1$. Then

$$L(t) = P(Y|MD, \tau = t) = 1_{\{v(t)=1\}},$$

and with N meioses (all with $Y_i = v_i(\tau) = 1$, $i = 1, \dots, N$) we obtain

$$L(t) = \prod_{i=1}^N 1_{\{v_i(t)=1\}} = 1_{\{v_1(t)=\dots=v_N(t)=1\}},$$

where $v_i(\cdot)$ is the inheritance process for meiosis i . Thus $\hat{\tau}_N = \arg \max_t L(t)$ is a union of finitely many intervals.

Asymptotics for $\hat{\tau}_N$:

Now $v(\cdot)|v(\tau) = 1$ is equiv. to two independent Markov processes evolving in either direction from τ with initial conditions $v(\tau) = 1$. We may write

$$\hat{\tau}_N = \arg \max_t Z_N(t) = \arg \max_t \frac{1}{N} \sum_{i=1}^N 1_{\{v_i(t)=1\}},$$

where

$$\begin{aligned} EZ_N(t) &= P(v_i(t) = 1) = 1 - \theta_{|t-\tau|} \\ &= \frac{1}{2} (1 + \exp(2\lambda|t - \tau|)). \end{aligned}$$

As $N \rightarrow \infty$,

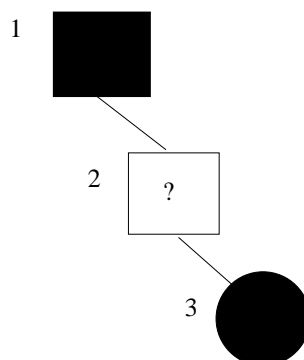
$$N(\hat{\tau}_N - \tau) \xrightarrow{\mathcal{L}} [T_{-1}, T_1],$$

where $-T_{-1}$ and T_1 are independent $\text{Exp}(\lambda^{-1})$.

Arbitrary setup

- Incomplete penetrance (arbitrary phenotypes)
 - Binary phenotypes
 - Quantitative phenotypes
 - Age-dependency/covariates
 - One- och multilocus models
 - Nonrandom mating and/or LD
- General pedigrees
 - Mixtures of different pedigrees in the same sample
- Arbitrary score functions
 - Likelihood
 - NPL
 - QTL

Incomplete penetrance



Let $Y = (Y_1, Y_3) = (1, 1)$ be the vector of observed binary phenotypes, where 1 = 'affected' and 0 = 'unaffected'. The *genetic parameters* for a single-locus model are

$$\begin{aligned} p &= P(A) = 1 - P(a) \\ f_0 &= P(Y_i = 1 | (aa)) \\ f_1 &= P(Y_i = 1 | (Aa)) \\ f_2 &= P(Y_i = 1 | (AA)) \end{aligned}$$

where A is the disease allele and a the normal allele. Here p is the *disease allele frequency* and $0 \leq f_0 \leq f_1 \leq f_2 \leq 1$ the *penetrance parameters*. Then $P(v(\tau) = 1 | Y) = q$, where usually $1/2 \leq q = q(p, f_0, f_1, f_2) \leq 1$. Notice that $q(p, f, f, f) = 1/2$ and $\lim_{p \rightarrow 0} q(p, 0, 1, 1) = 1$.

Asymptotics

$v(\cdot)|Y$ evolves as two independent Markov processes in either direction from τ , with initial conditions $P(v(\tau) = 1|Y) = q$. One can show that

$$C_1 \log L(t) + C_2 = \frac{1}{N} \sum_{i=1}^N 1_{\{v_i(t)=1\}} =: Z_N(t)$$

for some constants C_1 and C_2 . The ML-estimator $\hat{\tau}_N$ thus maximizes Z_N , and

$$N(\hat{\tau}_N - \tau) \xrightarrow{\mathcal{L}} \arg \max_{-\infty < s < \infty} \tilde{Z}(s),$$

where \tilde{Z} is a certain compound Poisson process, which can make jumps downwards *and* upwards if $1/2 \leq q < 1$.

Definition of \tilde{Z} :

The limiting compound Poisson process is defined by

$$\tilde{Z}(s) = \begin{cases} \sum_{0 < T_i \leq s} X_i, & s > 0, \\ \sum_{s \leq T_i < 0} X_i, & s < 0, \end{cases}$$

where $0 < T_1 < T_2 \dots$ and $0 > T_{-1} > T_{-2} \dots$ are two independent Poisson processes with intensity λ and $\{X_i\}$ are i.i.d., with

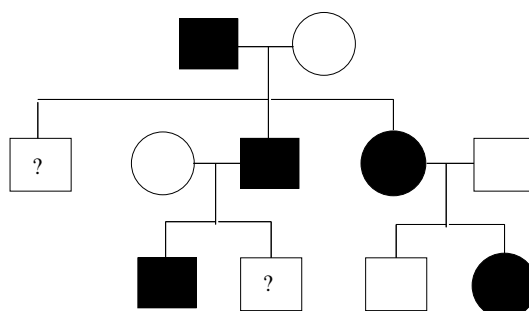
$$P(X_i = -1) = q, \quad P(X_i = 1) = 1 - q.$$

Notice that $\arg \max_s \tilde{Z}(s) = [T_{-1}, T_1]$ if $q = 1$.

The first two moments functions of \tilde{Z} are

$$\begin{aligned} E(\tilde{Z}(s)) &= -(2q - 1)\lambda|s|, \\ \text{Var}(\tilde{Z}(s)) &= \lambda|s|. \end{aligned}$$

General pedigree



For a general pedigree \mathcal{P} , we define

n = number of individuals (= 11)

f = number of founders (= 4)

m = number of meioses = $2(n - f)$ (= 14)

Y = set of known phenotypes = $\{Y_i, i \in \bar{\mathcal{P}}\}$
 where $\bar{\mathcal{P}} \subset \mathcal{P}$ has 9 individuals

$v(t)$ = $(v_1(t), \dots, v_m(t))$, inheritance vector of \mathcal{P}
 at locus t . $v_j(t) = 1$ if j :th meiosis transmits
 a paternal allele

$S(v(t))$ = $S(v(t); Y)$, a mapping $\mathbb{Z}_2^m \rightarrow \mathbb{R}$ measuring the
 compatibility between $v(t)$ and Y

Total score, N identical families

Suppose we have N families with the same pedigree structure and vector of known phenotypes. The total score is defined as

$$Z_N(t) = \frac{1}{N} \sum_{i=1}^N \bar{Z}_i(t),$$

where $\bar{Z}_i(t) = S(v_i(t); Y)$ is the *family score* and $v_i(t) = (v_{i1}(t), \dots, v_{im}(t))$ the inheritance vector of the i :th family.

In order to find the asymptotic behaviour of

$$\hat{\tau}_N = \arg \max_{0 \leq t \leq l} Z_N(t),$$

we must compute the first two moments of the i.i.d. terms $\bar{Z}_i(t)$.

Inheritance process, general case

A priori, the inheritance process $v(t) = (v_1(t), \dots, v_m(t))$, $0 \leq t \leq l$, consists of m independent Markov processes with jump intensity λ . Thus v has intensity matrix

$$A = \left(\begin{array}{cc} -\lambda & \lambda \\ \lambda & -\lambda \end{array} \right)^{m \otimes},$$

i.e. $A : \mathbb{Z}_2^m \times \mathbb{Z}_2^m \rightarrow \mathbb{R}$ is given by

$$A(w, w') = \begin{cases} -m\lambda, & w = w', \\ \lambda, & |w' - w| = 1, \\ 0, & |w' - w| \geq 2, \end{cases}$$

where $|w' - w| = \sum_{j=1}^m |w'_j - w_j|$ is the Hamming distance between w and w' .

Inheritance process given phenotypes

Let $P : \mathbb{Z}_2^m \rightarrow \mathbb{R}$ be a row vector defined by

$$P(w) = P(v(\tau) = w|Y),$$

i.e. the conditional distr. of the inheritance vector at the disease locus.

For $m = 1$ we had $P(v(\tau) = 1|Y) = q$, corresponding to $P = (P(0), P(1)) = (1 - q, q)$.

$v(\cdot)|Y$ evolves as two independent Markov processes in either direction from τ , with intensity matrix A . Thus

$$P(v(t) = w) = 'w\text{:th component of } P \exp(A|t - \tau|)'$$

Local behaviour of family scores

The formula for $P(v(t) = w)$ implies (after some computations) that

$$\begin{aligned} E(\bar{Z}(t)) - E(\bar{Z}(\tau)) &= -a|t - \tau| + o(|t - \tau|), \\ \text{Var}(\bar{Z}(t) - \bar{Z}(\tau)) &= \sigma^2|t - \tau| + o(|t - \tau|), \end{aligned}$$

as $t \rightarrow \tau$, where

$$\begin{aligned} a &= -SAP', \\ \sigma^2 &= SBS'. \end{aligned}$$

and $B = \text{diag}(PA) - \text{diag}(P)A - A\text{diag}(P)$. We refer to

$$\text{ASLNR} = \frac{a^2}{\sigma^2}$$

as the *asymptotic slope-to-noise ratio*. It will be crucial for determining estimation accuracy.

Asymptotics, N identical families

We look at $Z_N(t) = \sum_{i=1}^N \bar{Z}_i(t)/N$ on a *local chromosomal scale* around τ by introducing

$$\begin{aligned} \tilde{Z}_N(s) &= \frac{a}{\sigma^2} N \left(Z_N(\tau) - Z_N\left(\tau + \frac{\sigma^2}{a^2} N^{-1} s\right) \right) \\ &\xrightarrow{\mathcal{L}} \tilde{Z}(s) \end{aligned}$$

on $D(-\infty, \infty)$, where \tilde{Z} is a compound Poisson process¹, scaled so that

$$E\tilde{Z}(s) = -|s| \text{ and } \text{Var}(\tilde{Z}(s)) = |s|.$$

Further,

$$\text{ASLNR} \cdot N(\hat{\tau}_N - \tau) \xrightarrow{\mathcal{L}} \arg \max_s \tilde{Z}(s),$$

so ASLNR quantifies estimation accuracy.

¹ \tilde{Z} has jump intensity $m\lambda/\text{ASLNR}$ and i.i.d. jumps X_i satisfying $E(X_i) = -\text{ASLNR}/(m\lambda)$ and $E(X_i^2) = \text{ASLNR}/(m\lambda)$.

Optimal score function

An asymptotically optimal score function (in terms of estimation accuracy) can be obtained by maximizing

$$\text{ASLNR} = \frac{-SAP'}{SBS'}$$

w.r.t. S (analogous to minimizing asymptotic variance in ordinary \sqrt{N} asymptotics). The solution

$$S = PAB^{-1}$$

is different than the ML score function $S = \log P$, i.e. $S(w) = \log P(w)$, $\forall w \in \mathbb{Z}_2^m$. The maximal ASLNR;

$$I(\mathcal{P}) = \sup_S \text{ASLNR} = (PAB^{-1}AP)^{-1}$$

can be interpreted as a Fisher information for \mathcal{P} in terms of disease locus estimation under perfect marker information.

Mixture of different pedigree types

For a data set with N arbitrary (possibly different) pedigrees, we define the total score function

$$Z_N(t) = \frac{1}{N} \sum_{i=1}^N \gamma_i \bar{Z}_i(t),$$

where $\bar{Z}_i(t) = S(v_i(t); Y_i)$ is the family score for the i :th pedigree ($=\mathcal{P}_i$) and γ_i the corresponding weight. Y_i contains the observed phenotypes for \mathcal{P}_i .

The distr. of $\bar{Z}_i(\cdot)$ depends on the *pedigree type*

$$\phi_i = (\mathcal{P}_i, \bar{\mathcal{P}}_i, Y_i, \text{genetic model})$$

where $\bar{\mathcal{P}}_i \subset \mathcal{P}_i$ consists of individuals with known phenotypes in \mathcal{P}_i .

Asymptotics (including weighting)

Suppose that the weights are chosen from some weight function $\gamma(\cdot)$ according to

$$\gamma_i = \gamma(\phi_i)$$

and that the empirical distribution of all pedigree types $\{\phi_i\}_{i=1}^N$ converges weakly to some measure ν as $N \rightarrow \infty$ on the 'space of pedigree types'. Then

$$\frac{a^2}{\sigma^2} N(\hat{\tau}_N - \tau) \xrightarrow{\mathcal{L}} \arg \max_s \tilde{Z}(s),$$

where \tilde{Z} is a compound Poisson process with $E\tilde{Z}(s) = -|s|$ and $\text{Var}(\tilde{Z}(s)) = |s|$. Further,

$$\begin{aligned} a &= \int \gamma(\phi) a(\phi) d\nu(\phi), \\ \sigma^2 &= \int \gamma^2(\phi) \sigma^2(\phi) d\nu(\phi) \end{aligned}$$

where $a(\phi)$ and $\sigma^2(\phi)$ are the mean slope and local incremental variance for a pedigree of type ϕ .

Optimal weighting

Optimal weighting can be derived by maximizing (given a certain score function)

$$\text{ASLNR}(\nu) = \frac{a^2}{\sigma^2} = \frac{(\int \gamma(\phi)a(\phi)d\nu(\phi))^2}{\int \gamma^2(\phi)\sigma^2(\phi)d\nu(\phi)}$$

w.r.t. $\gamma(\cdot)$. Cauchy-Schwarz' inequality gives the optimal solution

$$\gamma(\phi) \propto a(\phi)/\sigma^2(\phi).$$

independently of the population measure ν !! Notice that negative $a(\cdot)$ and $\gamma(\cdot)$ are allowed for. Optimally, $a(\cdot)$ and $\gamma(\cdot)$ should have the same sign.

Other Applications

- **Sampling:** Given a pedigree \mathcal{P} with observed phenotypes Y , we can use $ASLNR(\delta_\phi)$ as a performance measure of the pedigree in terms of possible estimation accuracy. Here ϕ is the type associated with (\mathcal{P}, Y) and the given genetic model.
- **Planning of marker maps:** Since $\hat{\tau}_N - \tau = O_p(ASLNR(\nu)^{-1}N^{-1})$ a map of markers with grid size δ is sufficient, with $\delta \sim ASLNR(\nu)^{-1}N^{-1}$.
- **Robustness w.r.t. choice of map function:** The asymptotic results are valid for a large class of map functions (= models for occurrence of crossovers), not only Haldane's Poisson model.

More general convergence rates

Assume there exist numbers $a, \sigma^2 > 0$ and $1/2 \leq \beta < \alpha$, $\beta \leq 1$ such that

$$\begin{aligned} E(Z_N(t)) - E(Z_N(\tau)) &= -a|t - \tau|^\alpha + o(|t - \tau|^\alpha), \\ \text{Var}(Z_N(t) - Z_N(\tau)) &= \sigma^2|t - \tau|^{2\beta} + o(|t - \tau|^{2\beta}). \end{aligned}$$

as $t \rightarrow \tau$. Then, under mild additional regularity conditions, $\hat{\tau}_N - \tau = O_p(N^{-d})$, where $d = 1/(2(\alpha - \beta))$. More precisely,

$$\left(\frac{a}{\sigma}\right)^{2d} N^d (\hat{\tau}_N - \tau) \xrightarrow{\mathcal{L}} \arg \max_s \tilde{Z}(s),$$

where \tilde{Z} satisfies

$$E(\tilde{Z}(s)) = -|s|^\alpha \text{ and } \text{Var}(\tilde{Z}(s)) = |s|^{2\beta}.$$

A typical example of limiting process is $\tilde{Z}(s) = B_\beta(s) - |s|^\alpha$, where B_β is a FBM with self-similarity parameter β .

Examples

Estimator	α	β	d	\tilde{Z}
ML, M	2	1	1/2	$B_1(s) - s^2$
M (jump scores)	2	1/2	1/3	$B_{1/2}(s) - s^2$
Linkage	1	1/2	1	Comp. Poisson

- B_β is a two-sided Fractional Brownian Motion with self similarity parameter β . Hence $B_{1/2}$ is an ordinary two-sided Brownian motion.
- Since $B_1(s) - s^2 = B_1(1)s - s^2$ with $B(1) \sim N(0, 1)$, $\arg \max_s (B_1(1)s - s^2) = B_1(1)/2$ is normally distributed.
- In the limit ASLNR $\rightarrow 0$, the compound Poisson process in linkage application approaches $B_{1/2}(s) - |s|$, the Gaussian process having the same first two moment functions as \tilde{Z} .