W05 - Maximum Likelihood - least squares - linear regression

Sivia ch 3, 5 ..

Lander and Botstein 1988

biol problem: Ding et al. 2015

Maximum Likelihood (ML)

D - data
H - hypothesis
X - parameters of H

$P(D|XH)$

$P(X|DH) = \frac{P(D|XH)\,P(X|H)}{P(D)}$

posterior:

$P(X|D,H) \propto P(D|XH) \cdot P(X|H)$

if we assume an uninformative prior $P(X|H) = \text{constant}$

$$\underset{X}{\text{argmax}}\, P(X|DH) = \underset{X}{\text{argmax}}\, P(D|XH)$$

$$= \underset{X}{\text{argmax}}\, \log P(D|XH)$$

X^* ML fit data is also X at which

posterior of X given D is also maximal.

(uninformative prior).

$L(X) = \log P(D|XH)$, $\left.\frac{\delta L}{\delta X}\right|_{X^*} = 0$

$$L(x) = \log P(D|XH)$$

$$\left.\frac{\delta L(x)}{\delta x}\right|_{x^*} = 0 \qquad \left.\frac{\delta^2 L}{\delta x^2}\right|_{x^*} = -\frac{1}{\sigma_2^*}$$

Distribution	$P(D\|XH)$	x^*	σ^*
exponential $D=\{t_1..t_N\}$	$\frac{1}{\lambda}e^{-t/\lambda}$	$\frac{e^{-\Sigma_i t_i/\lambda}}{\lambda^N}$	$x^* = \frac{\Sigma_i t_i}{N}$	$\lambda^*/\sqrt{N}$
binomial (P) n, N	P $(1-P)$	$\binom{N}{n}p^n(1-p)^N$	$P^* = \frac{n}{N}$	$\frac{\sqrt{P^*(1-P^*)}}{\sqrt{N}}$
Gaussian μ,σ $D=\{t_1..t_N\}$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N e^{-\frac{\Sigma_i (t_i-\mu)^2}{2\sigma^2}}$	$\mu^* = \frac{\Sigma_i t_i}{N}$ $\sigma^* = \sqrt{\frac{\Sigma_i (t_i-\mu)^2}{N}}$	$\sigma_\mu^* = \sigma^*/\sqrt{N}$ $\sigma_\sigma^* = \sigma^*/\sqrt{2N}$

$$P(D|\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N e^{-\frac{\sum_i (t_i-\mu)^2}{2\sigma^2}}$$

$$\log P(D|\mu,\sigma) = -N\log\sqrt{2\pi} - N\log\sigma - \frac{\sum_i (t_i-\mu)^2}{2\sigma^2}$$

$$\frac{\delta L(\mu,\sigma)}{\delta \mu} = -\frac{1}{2\sigma^2}\cdot 2\sum_i (t_i-\mu)\cdot(-1) = +\frac{1}{\sigma^2}\sum_i (t_i-\mu)$$

$$\frac{\delta L(\mu,\sigma)}{\delta \sigma} = -\frac{N}{\sigma} - (-2)\frac{\sum_i (t_i-\mu)^2}{2\sigma^3} = -\frac{1}{\sigma}\left[N - \frac{\sum_i (t_i-\mu)^2}{\sigma^2}\right]$$

$$\left.\frac{\delta L}{\delta \mu}\right| = 0 \quad \Rightarrow \quad \boxed{\mu^* = \frac{1}{N}\sum_i t_i}$$

$$\left.\frac{\delta L}{\delta \sigma}\right| = 0 \quad \Rightarrow \quad \boxed{\sigma^{*2} = \frac{\sum_i (t_i-\mu)^2}{N}}$$

$$\frac{\delta^2 L}{\delta \mu^2} = -\frac{N}{\sigma^2} \qquad \Big|_{\mu^*,\sigma^*} = -\frac{N}{\sigma_*^2} \Rightarrow \boxed{\sigma_\mu^{*2} = \frac{\sigma_*^2}{N}}$$

$$\frac{\delta^2 L}{\delta \sigma^2} = +\frac{N}{\sigma^2} - 3\frac{\sum_i (t_i-\mu)^2}{\sigma^4}\Big|_{\mu^*,\sigma^*} = \frac{N}{\sigma_*^2} - 3\frac{N\sigma_*^2}{\sigma_*^4}$$

$$= -2\frac{N}{\sigma_*^2} \Rightarrow \boxed{\sigma_\sigma^{*2} = \frac{\sigma_*^2}{2N}}$$

Least-squares fit

$$D = \{d_i\}_{i=1}^{N}$$

$$P(D|XH) = \prod_{i=1}^{N} P(d_i|XH)$$

Assume we know how ideal data behaves except for a noise

$$d_i = f_i(x)$$

$$d_i - f_i(x) \sim W(0, \sigma_i)$$ Gaussian noise
- symmetric
- not too large

$$\sigma_i \simeq \sigma$$

$$P(D|XH) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{N} e^{-\frac{\sum_i (d_i - f_i(x))^2}{2\sigma^2}}$$

$$\log P(D|XH) \propto -N \log \sigma - \frac{\sum_i (d_i - f_0(x))^2}{2\sigma^2}$$

$$\chi^2 = \sum_i (d_i - f_i(x))^2$$ least

$$\arg\max_X \log P(D|XD) = \arg\max_X \chi^2$$

least square fit

L^2-norm.

Particular case fit to a line

$d_i = \{x_i, y_i\}_{i=1}^{N}$

$$y_i - (a + b x_i) \sim N(0, \sigma)$$

$$P(D|a,b,H) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N e^{-\frac{\sum_i (y_i - a - b x_i)^2}{2\sigma^2}}$$

$$\underset{a,b}{\text{argmax}}\, P(D|a,b,H) = \underset{a,b}{\text{argmax}}\, \sum_i (y_i - (a+b x_i))^2$$

linear regression

QTL analysis

crossable S_1, S_2 > have a phenotypic difference

Which genomic region is responsible for it?

noisy: many loci
large environment contribution.

To: identify the loci responsible for a particular trait.

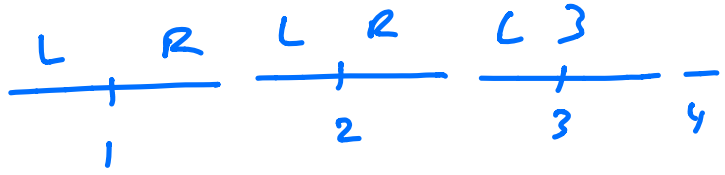
Ding et al 2016

× (Drosophila Simulans (Dsim) $\mu_{sim} \simeq 176$ Hz
Drosophila mauritiana (Dmau) $\mu_{mau} = 186$ Hz

male courtship sine song

Where does the difference come from?

Dsim, Dmau ~ 98% similarity of genomes

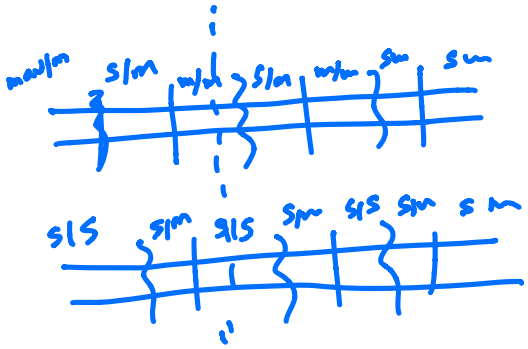


Sim/Sim ═══ Dsim ♂ × Dmau ♂ ═══ mau/mau

↓

F_1 ═══ sim/mau

Backcrosses



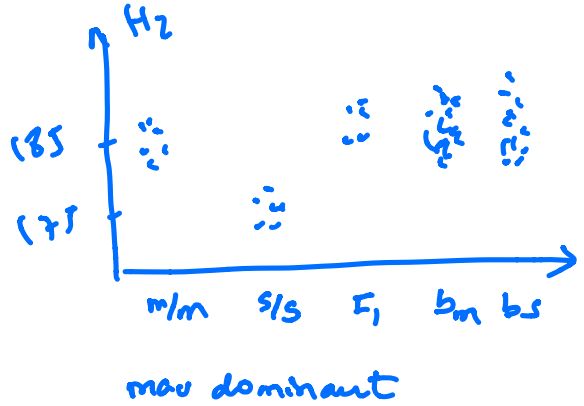
~1 breakpoint across the

From backcrosses collect in female progeny i'

+ phenotype $\{\mu_i\}_{i=1}^{L}$

+ genotype $\{g_i\}_{i=1}^{L}$

$$g_i = \begin{cases} 1 & \text{if } m/m \\ & m/s \\ 0 & s/s \end{cases}$$



notice: how to get the haplotype assignment is another computational question in itself that we will study in workflow using hidden Markov models (HMMs)

The models — assume one locus

+ QTL model

$$f_i = a + b\,g_i + \epsilon_i$$

(f_i: phenotype; g_i: genotype)

$$P(D \mid a, b, \mathrm{QTL}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^N e^{-\frac{\sum_i (f_i - a - b g_i)^2}{2\sigma^2}}$$

+ NQTL model

$$f_i = c + \epsilon_i$$

$$P(D \mid c, \mathrm{NQTL}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^N e^{-\frac{\sum_i (f_i - c)^2}{2\sigma^2}}$$

QTL → 2 parameters a, b

NQTL → 1 " c

(if $a = c$ and $b = 0$ identical)

$$\frac{P(\mathrm{QTL} \mid D)}{P(\mathrm{NQTL} \mid D)} = \frac{P(D \mid \mathrm{QTL}) \cdot P(\mathrm{QTL})}{P(D \mid \mathrm{NQTL}) \cdot P(\mathrm{QTL})} = \frac{P(D \mid \mathrm{QTL})}{P(D \mid \mathrm{NQTL})}$$

equal priors

$$P(D \mid \mathrm{QTL}) = \int_{-\infty}^{+\infty} da \int_{-\infty}^{+\infty} db\; P(D \mid a, b)\; P(a, b \mid \mathrm{QTL})$$

$$P(D \mid \mathrm{NQTL}) = \int_{-\infty}^{+\infty} dc\; P(D \mid c)\; P(c \mid \mathrm{NQTL})$$

priors for parameters

$$P(a,b \mid QTL) = P(a \mid QTL) \cdot P(b \mid QTL)$$

$$P(a \mid QTL) = \begin{cases} \frac{1}{a^+ - a^-} = \frac{1}{\sigma_a} & a^- < a < a^+ \\ 0 & \text{otherwise} \end{cases}$$

$$P(b \mid QTL) = \begin{cases} \frac{1}{b^+ - b^-} = \frac{1}{\sigma_b} & b^- < b < b^+ \\ 0 & \text{otherwise} \end{cases}$$

$$P(c \mid NQTL) = \begin{cases} \frac{1}{c^+ - c^-} = \frac{1}{\sigma_c} & c^- < c < c^+ \\ 0 & \text{otherwise} \end{cases}$$

Laplace approximation: for $P(D|a,b,Q,T,N)$

$P(D|c,N,Q,T,L)$

$$L^Q(a,b) =: \log P(D|a,b,Q,T,L)$$

$$L^N(c) =: \log P(D|c,N,Q,T,L)$$

Taylor expansion around ML values a^*, b^*, c^*

$$L^Q(a,b) \simeq L^Q(a^*,b^*) + \frac{1}{2}\left.\frac{\delta^2 L^Q}{\delta a^2}\right|_* (a-a^*)^2$$

$$\left.\frac{\delta L^Q}{\delta a}\right|_{a^*,b^*} = 0 \qquad + \frac{1}{2}\left.\frac{\delta^2 L^Q}{\delta b^2}\right|_* (b-b^*)^2$$

$$\left.\frac{\delta L^Q}{\delta b}\right|_{a^*,b^*} = 0 \qquad + \left.\frac{\delta^2 L}{\delta a \delta b}\right|_* (a-a^*)(b-b^*)$$

$$L^N(c) \simeq L^N(c^*) + \frac{1}{2}\left.\frac{\delta^2 L^N}{\delta c^2}\right|_* (c-c^*)^2$$

$$\left.\frac{\delta L^N}{\delta c}\right|_{c^*} = 0$$

ML values of QTL / NQTL parameters

$$L^Q \propto -\frac{\sum_i (f_i - a - b g_i)^2}{2\sigma^2}$$

$$L^N \propto -\frac{\sum_i (f_i - c)^2}{2\sigma^2}$$

$$\frac{\delta L^Q}{\delta a} = -\frac{2}{2\sigma^2}\sum_i (f_i - a - b g_i)(-1) = \frac{1}{\sigma^2}\sum_i (f_i - a - b g_i)$$

$$\frac{\delta L^Q}{\delta b} = -\frac{2}{2\sigma^2}\sum_i (f_i - a - b g_i)(-g_i) = \frac{1}{\sigma^2}\sum_i (f_i - a - b g_i) g_i$$

$$= \frac{1}{\sigma^2}\left(\sum_i f_i g_i - a\sum_i g_i - b\sum_i g_i^2\right)$$

$$\bar{f} = \frac{1}{N}\sum_i f_i \qquad \overline{f^2} = \frac{1}{N}\sum_i f_i^2$$

$$\bar{g} = \frac{1}{N}\sum_i g_i \qquad \overline{g^2} = \overline{gg} = \frac{1}{N}\sum_i g_i^2$$

$$\overline{gf} = \overline{fg} = \frac{1}{N}\sum_i f_i g_i$$

$$\frac{\delta L^Q}{\delta a} = \frac{N}{\sigma^2}\left(\bar{f} - a - b\bar{g}\right) \Big| = 0 \quad \bar{f} = a^* + b^*\bar{g}$$

$$\frac{\delta L^Q}{\delta b} = \frac{N}{\sigma^2}\left(\overline{fg} - a\bar{g} - b\overline{gg}\right) \Big| = 0 \quad \overline{fg} = a^*\bar{g} + b^*\overline{gg}$$

$$a^* = \bar{f} - b^*\bar{g}$$

$$\overline{fg} = (\bar{f} - b^*\bar{g})\bar{g} + b^*\overline{gg}$$

$$\frac{\delta L^N}{\delta c} = \frac{N}{\sigma^2}(\bar{f} - c)\Big|_0 = 0 \Rightarrow \boxed{c^* = \bar{f}}$$

$$b^*\left[\overline{gg} - \bar{g}^2\right] = \overline{fg} - \bar{f}\bar{g}$$

ML

QTL

$$b^* = \frac{\overline{fg} - \bar{f}\bar{g}}{\overline{g^2} - \bar{g}^2}$$

$$a^* = \bar{f} - b^* \bar{g}$$

NQTL

$$c^* = \bar{f}$$

$$\frac{\delta L^a}{\delta a} = \frac{N}{\sigma^2}(\bar{f} - a - b\bar{g})$$

$$\frac{\delta L^a}{\delta b} = \frac{N}{\sigma^2}(\overline{fg} - a\bar{g} - b\overline{g^2})$$

$$\frac{\delta L^N}{\delta c} = \frac{N}{\sigma^2}(\bar{f} - c)$$

$$\frac{\delta^2 L^a}{\delta a^2} = -\frac{N}{\sigma^2} \qquad \frac{\delta^2 L^a}{\delta a \delta b} = -\frac{N}{\sigma^2}\bar{g}$$

$$\frac{\delta^2 L^a}{\delta b^2} = -\frac{N}{\sigma^2}\overline{g^2}$$

$$\frac{\delta^2 L^N}{\delta c^2} = -\frac{N}{\sigma^2}$$

$$P(D|a\,b\,Q\,T_L) \simeq P(D|a^*b^*Q\,T_L) \times \exp\left[-\frac{N}{2\sigma^2}(a-a^*)^2 - \frac{N}{2\sigma^2}\bar{g_g}(b-b^*)^2 - \frac{N}{\sigma^2}\bar{f}(a-a^*)(b-b^*)\right]$$

$$P(D|c\,N\,Q\,T_L) \simeq P(D|c^*\,N\,Q\,T_L) \exp\left[-\frac{1}{2}\frac{N}{\sigma^2}(c-c^*)^2\right]$$

$$P(D|a\,b\,Q\,T_L) \simeq P(D|a^*b^*Q\,T_L) \exp\left[\begin{pmatrix}a-a^* & b-b^*\end{pmatrix}\begin{bmatrix}1 & \bar{f} \\ \bar{f} & \bar{g_g}\end{bmatrix}\begin{pmatrix}a-a^* \\ b-b^*\end{pmatrix}\frac{N}{\sigma^2}\right]$$

$$P(D|Q\,T_L) = P(D|a^*b^*Q\,T_L)\cdot\int_{-\infty}^{+\infty}da\int_{-\infty}^{+\infty}db\; e^{-\frac{1}{2}(a-a^*,b-b^*)\,Q\begin{pmatrix}a-a^* \\ b-b^*\end{pmatrix}} \times\frac{1}{\sigma_a}\frac{1}{\sigma_b}$$

$$Q = \frac{N}{\sigma^2}\begin{bmatrix}1 & \bar{f} \\ \bar{f} & \bar{g_g}\end{bmatrix} \simeq P(D|a^*b^*Q\,T_L)\frac{1}{\sigma_a}\frac{1}{\sigma_b}\frac{2\pi}{\sqrt{\det Q}}$$

$$\det Q = \left(\frac{N}{\sigma^2}\right)^2(\bar{g_g}-\bar{f}^2) \simeq P(D|a^*b^*Q\,T_L)\frac{2\pi\sigma^2}{\sigma_a\sigma_b}\frac{1/N}{\sqrt{\bar{g_g}-\bar{f}^2}}$$

$$\simeq P(D|a^*b^*Q\,T_L)\frac{2\pi/N}{\sqrt{\bar{g_g}-\bar{f}^2}}\cdot\frac{\sigma}{\sigma_a}\frac{\sigma}{\sigma_b}$$

$$P(D|c\,N\,Q\,T_L) \simeq P(D|c^*N\,Q\,T_L)\frac{1}{\sigma_c}\int_{-\infty}^{+\infty}e^{-\frac{1}{2}\frac{N}{\sigma^2}(c-c^*)^2}dc$$

$$\simeq P(D|c^*N\,Q\,T_L)\frac{\sqrt{2\pi}\,\sigma/\sqrt{N}}{\sigma_c}$$

$$\frac{P(D|Q\,T_L)}{P(D|N\,Q\,T_L)} = \underbrace{\frac{P(D|a^*b^*Q\,T_L)}{P(D|c^*N\,Q\,T_L)}}_{n_L}\underbrace{\sqrt{\frac{2\pi/N}{\bar{g_g}-\bar{f}^2}}\frac{\sigma/\sigma_a\;\sigma/\sigma_b}{\sigma/\sigma_c}}_{\text{Occam's razor}}$$

LOD scores

Ding et al were not bayesian's, but they used a method similar to Lander's.

$$LOD = \log \frac{P(D \mid a^i b^j \theta_{TL})}{P(D \mid c^k N \theta_{TL})}$$

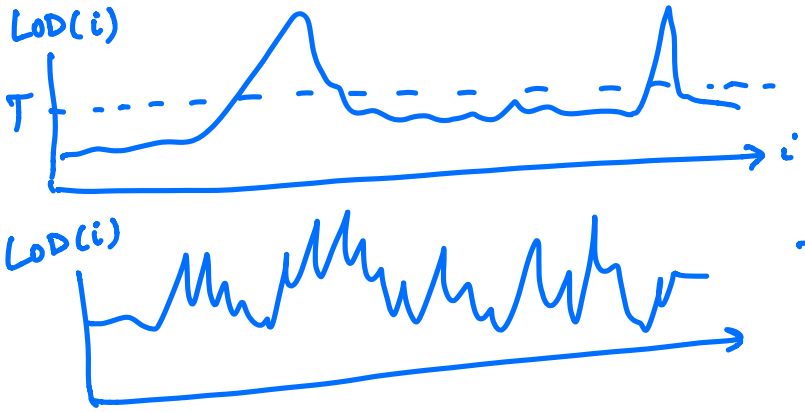
$$= \log P(D \mid a^i b^j \theta_{TL}) - \log P(D \mid c^k N \theta_{TL}) \geq 0$$

nested hypotheses

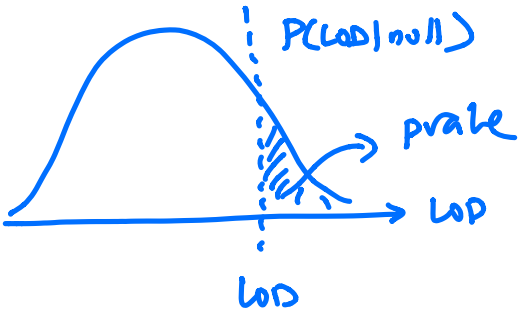
"Important issue"

" What LOD threshold T should be used in order to maintain an acceptable low positive rate? "

Lander Fig 2.3



+ Take a collection of pairs (f_i, s_i) known to be non-associated $\sim H_0$



expected #FP $\simeq$ $P(LOD \geq LOD(i) \mid null) \times N$

N = number of loci tested

Generalization

1 phenotype – 1 loci $f_i = a + b g_i$

1 phenotype – κ loci

$$f_i = a + \sum_{\kappa=1}^{\kappa} b^{\kappa} g_i^{\kappa} + \epsilon_i$$

Assumptions

1- linear fit btw $f_i \sim g_i^{\kappa}$

2- Gaussian noise

3- all loci have same noise
all individuals

4- σ is known.