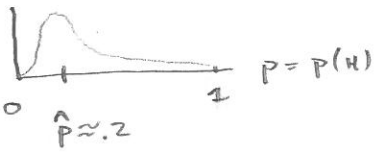
Model-based Inference

In statistics, there are two frameworks for model-based inference:

Maximum Likelihood	Bayesian Analysis
"frequentists"	"Bayesians"
$p(H) = p$ unknown	get a POSTERIOR DIST. on your parameter of interest
HTHTTTH ----> ∞	
$\hat{p}$ = proportion of H → p in n trials	

We begin with the Likelihood framework. Warning: Need to review optimization

Easy example: Show that the Maximum Likelihood Estimator (MLE) $\hat{p}$ for n coin flips is $\hat{p} = \frac{\#\text{ of H}}{n} \equiv$ proportion of H in n tosses

Setup: 1) $p = p(H)$ is unknown and will be estimated

2) there are n coin flips and n_H are H, n_T are T, $n_H + n_T = n$.

The data are $n_H, n_T \equiv$ data,

With p unknown, define the likelihood function

(Assumes i.i.d.)

$$L(p) = \text{Prob}\left(\text{data} \,\middle|\, p\right) = p^{n_H}(1-p)^{n_T}$$

then the MAXIMUM LIKELIHOOD ESTIMATOR $\hat{p} = \hat{p}_{MLE}$ is the value of p that maximizes $L(p)$. [The value of p that makes the data you observed most probable.]

Eg. $n = 100 \quad n_H = 30 \quad n_T = 70$

$$L(p) = \text{Prob}(\text{data} \mid p) = p^{30}(1-p)^{70}$$

Quick Calc I optimization:

$$L'(p) = p^{30}\left[70(1-p)^{69}(-1)\right] + 30p^{29}(1-p)^{70}$$

$$= p^{29}(1-p)^{69}\left[30(1-p) - 70p\right]$$

Require $L'(p) = 0 \rightarrow$

$$p^{29}(1-p)^{69}\left[30 - 100p\right] = 0 \quad \text{i.e.} \quad p = 0, 1, \frac{30}{100} = .3$$

Using your favorite test, it is clear $p = .3$ is a global max.

$$\therefore \quad \hat{p} = \hat{p}_{MLE} = .3 = \frac{n_H}{n} \qquad \square$$

Redo: Since ML will require set derivatives to zero (and typically iid assumption) in practice, use the Log-likelihood function.

Since $\log(x)$ is an increasing function, the log-likelihood and likelihood function are maximized at the same value of p.

$$\text{log-likelihood} \equiv \ln L(p) = \ln\left(p^{30}(1-p)^{70}\right) = \ln p^{30} + \ln(1-p)^{70}$$

Differentiate: $\frac{d}{dp}\ln L(p) = \frac{1}{p^{30}} - \frac{1}{(1-p)^{70}}$ Easy to differentiate...

has critical point at $p = 0, 1, \doteq .3$ ...

Caution: MLEs may fail to exist or not be unique.

Maximum Likelihood Trees in phylogenetics

Goal: Fix a site substitution model JC, K2P, GTR

and find the parameters $T, (\vec{p}, Q, \{t_e\})$ that maximize the likelihood

Example: Model Jukes-Cantor on 1-edge tree 

$\in T$

$\vec{p} = (.25\ .25\ .25\ .25)$ known

$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ & \text{etc.} & & \end{pmatrix}$ known

Unknown parameter t = branch length

Data: F_{ab} = pattern frequencies counts of pattern (under iid assumption)

$$= \begin{pmatrix} n_{AA} & n_{AG} & n_{AC} & n_{AT} \\ n_{GA} & & \ddots & \\ n_{CA} & & & \\ n_{TA} & & & n_{TT} \end{pmatrix}$$ 16 bits of data n_{ij}

Under JC, the expected pattern frequency array is $\begin{pmatrix} 1-a(t) & a(t) & a(t) & a(t) \\ \overline{4} & \overline{12} & \overline{12} & \overline{12} \\ & & \ddots & \end{pmatrix}$

Where $a(t)$ is a invertible function of the unknown parameter t

$$a(t) = a = \frac{3}{4}\left(1 - e^{-\frac{4}{3}t}\right)$$

Thus, the log-likelihood is

$$\ln L(a) = \ln L(a \mid \text{data}) = \ln\left(\text{Prob}(\text{data} \mid a)\right)$$

$$= \ln\left[p_{AA}^{n_{AA}} p_{AG}^{n_{AG}} \cdots p_{TT}^{n_{TT}} \right] = \ln\left(\prod_{\substack{\text{pattern} \\ ij}} p_{ij}^{n_{ij}} \right)$$

$$= \ln\left[p_{ii}^{n_{AA}+n_{GG}+n_{CC}+n_{TT}} \cdot p_{ij}^{\sum_{i,j=1,\, i \neq j}^{4} n_{ij}} \right]$$

$$\ln L(a) = \ln\left[\left(\frac{1-a}{3}\right)^{\sum n_{ii}} \left(\frac{a}{12}\right)^{\sum_{i\neq j} n_{ij}}\right] = \left(\sum n_{ii}\right)\ln\left(\frac{1-a}{3}\right) + \left(\sum n_{ij}\right)\ln\left(\frac{a}{12}\right)$$

Differentiating ...

$$\frac{d}{dt}\ln L(a(t)) = \sum n_{ii}\left(\frac{3}{1-a}\right)\left(\frac{-1}{3}\right).a'(t) + \sum_{i\neq j} n_{ij}\,\frac{12}{a}\left(\frac{1}{12}\right)a'(t)$$

$$= -\sum n_{ii}\,\frac{1}{1-a(t)}\,a'(t) + \sum_{i\neq j} n_{ij}\,\frac{1}{a}\,a'(t)$$

Assume $a'(t) \neq 0$ (i.e. there is change!) and require equal to zero

$$\frac{d}{dt}\ln L(a(t)) = 0 \Rightarrow \sum n_{ii}\,\frac{1}{1-a} = \sum n_{ij}\,\frac{1}{a}$$

$$\Rightarrow \left(\sum n_{ii}\right)a = \left(\sum n_{ij}\right) - \left(\sum n_{ij}\right)a \Rightarrow a = \frac{\sum n_{ij}}{\sum n_{ii} + \sum n_{ij}}$$

i.e. $\hat{a}(t) = \dfrac{\sum n_{ij}}{n}$ = proportion of sites that differ!

and $\hat{t} = \hat{t}_{MLE} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{a}\right) = d_{JC}(a,b)$

i.e. The JC distance is the Maximum Likelihood estimate for t!

$$\hat{t}_{MLE} = d_{JC}(a,b)$$

Again justifies use of d_{JC} or corrected distances more generally.

Inferring an ML tree

1) Choose a model

Say: GTR

parameters: T unrooted since time-reversible

$\vec{p} = (p_A \; p_G \; p_C \; p_T)$ 3

$Q = Q(\alpha, \beta, \gamma, \delta, \epsilon, \eta)$ 5 Assume $\eta = 1$

$= ($ $)$

branch lengths $\{t_e\}$ $2n-3$ binary

$t_e \geq 0$

Total: $3+5+2n-3 = 2n+5$

2) Assuming characters (sites) are generated iid and using notation $p_{i_1 \cdots i_n}$ for $P(s_1 = i_1, \ldots, s_n = i_n)$

i.e. expected value of pattern $i_1 i_2 \cdots i_n$,

then the log-likelihood is

$$\ln L_T(\vec{p}, Q, \{t_e\}) = \ln \prod_{\substack{\text{all patterns} \\ i_1 i_2 \cdots i_n}} p_{i_1 \cdots i_n}^{n_{i_1 \cdots i_n}}$$

where $n_{i_1} \cdots n_{i_n}$ is the count of pattern $i_1 \cdots i_n$ in the sequence data.

$$= \sum_{\substack{\text{all pattern} \\ i_1 \cdots i_n}} n_{i_1 \cdots i_n} \ln(p_{i_1 \cdots i_n}) = \text{Prob}(\text{data} \mid p_A, p_G, p_C, \alpha, \beta, \gamma, \delta, \epsilon, \{t_e\})$$

3) For T fixed, optimize $\ln L_T$ to find MLE on T (i.e. maximizers $\hat{\vec{p}}, \hat{Q}, \{\hat{t_e}\}$ on T.

Save maximal value to (largest log-likelihood on T)

4) Repeat step 3 for all $(2n-5)!!$ other trees

5) Return tree T and parameters with largest log-likelihood of all iterations.

Issues/ Observations

1) Must find the MLE for all $(2n-5)!!$ trees, tree space is huge!

NP-hard

2) Maybe there is not a unique MLE ...

3) Optimization techniques for a fixed T are well-developed. (Felsenstein pruning algorithm)

4) ML is an attractive method since

- its well grounded in Statistics
- model assumptions are explicit

Eg: is a stable base dist. reasonable?

common Q?