

Lecture notes on Bryant and Waddell Paper

Prepared by Jun Han

1. The representation of taxon distance, edge length and topology of phylogenetic tree.

Let N be the number of taxa in the set L of taxa, n be the number of edges in the tree T , m be the number of pairs of distances among taxa. The taxa distances and edge length are related as

$$P = Ab$$

where P ($m \times 1$ vector) is taxon-to-taxon distances, b ($n \times 1$ vector) is the edge lengths and A ($m \times n$ matrix) is the topological matrix for the tree T . Here $m = N(N - 1)/2$.

Example: Refer to Bryant and Waddell paper Figure 1(i) (p1349), regard the four clusters C_j, C_l, C_m, C_k as four taxa j, l, m, k and assign the edge lengths as $e_j = 2, e_i = 4, e_l = 6, e_m = 3, e_k = 1$. Then $L = [j, l, m, k]$, $N = 4, m = 6, n = 5$ and $P = Ab$ is as follows:

$$\begin{bmatrix} P_{jl} = 12 \\ P_{jm} = 9 \\ P_{jk} = 3 \\ P_{lm} = 9 \\ P_{lk} = 11 \\ P_{mk} = 8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_j = 2 \\ b_i = 4 \\ b_l = 6 \\ b_m = 3 \\ b_k = 1 \end{bmatrix}$$

2. The mathematical formulation of least square method

Given the taxon-to-taxon distance d and topological matrix A of the phylogenetic tree, we want to find the edge length b that minimizes the criteria

$$Q(b) = (Ab - d)^T (Ab - d) = \|Ab - d\|_2^2$$

which is a Quadratic form (in vector case) or a square (in scalar case like the above 2-norm). Since we want to minimize the above square to be the least, so it is called least square method.

3. The statistical regression approach using least square.

One of the statistical approaches to evaluate the phylogenetic tree is to regard the taxon-to-taxon distance d as random vector and appeal to multiple regression. The regression model for the phylogenetic tree is

$$d = Ab + e$$

where d is an observed distance, A is the relation among species (topological matrix) and $e \sim N(0, \sigma^2 I)$ is random error. In this model, the expectation of d is $E[d] = Ab$ and the variance of d is $var[d] = V$ assumed to be $\sigma^2 I$. In other words, the distances are assumed to have constant variances and no correlations. This is the flavour of Ordinary Least Square (OLS), since ordinary least square method is applied to obtain the estimate of b in the above regression model under the above assumptions.

Since it is reasonable to assume that the distances have their own variances and are correlated each other, so the above assumptions are not realistic. Note that all the statistical inferences are only valid under the assumptions of OLS. To solve this dilemma, we apply OLS to $V^{\frac{1}{2}}d$ since $V^{\frac{1}{2}}d$ have constant variances and are uncorrelated. Now we get both the valid statistical inferences and realistic assumptions. This method is called Generalized Least Square (GLS).

The normal equation solution by GLS is given by

$$\hat{b} = (A^T V^{-1} A)^{-1} A^T V^{-1} d$$

where \hat{b} is the GLS estimate of edge length b . If V is a full SPD (symmetric positive definite) matrix, the solution is GLS solution; if $V = W^{-1}$ is positive diagonal matrix, the solution is WLS (Weighted Least Square) solution; if $V = \sigma^2 I$ is positive constant diagonal matrix, the solution is called OLS solution. Of course, the GLS solution is the best, other two are special cases of GLS.