Tree uncertainty...: Supplementary Materials

Amy Willis and Rayna Bell

Reducing computation when projecting

We consider incorporating intrinsic tree variability information, because the extrinsic case is a simplification (computationally). The setting is that we have the "reduced" dataset $Y_i \in \mathbb{R}^{n_i \times m}$, with row k given by $\Phi_{\overline{T}}(\hat{T}_k^{(i)})$, $k = 1, \ldots, n_i$, and the full data set $X = (Y_1^T, \ldots, Y_I^T)^T$, which just combines the reduced datasets. We want to project the Y_i onto the first 2 principal components of X. We refer to Y_i as Y, and merely repeat the procedure for all i.

Let $U\Sigma V^T$ bv singular value decomposition of X^T. We know that the projection matrix for projecting onto the first two principal components of X is given by $R = (U\Sigma^{-1})_{[,1:2]}$.

The minimum volume set containing measure $(1-\alpha)$ of a $\mathcal{N}(\bar{y}, \hat{\Sigma}_y)$ distribution is parameterised by

$$\begin{aligned} E &= \{z | (z-\bar{y})\Sigma^{-1}(z-\bar{y}) < mF_{m,n_i-1}(1-\alpha)\} \\ &= \{z | z = \bar{y} + L_y^{-T}u \text{ for } u \in B_m\}, \end{aligned} \tag{1}$$

where $B_m = \{x \in \mathbb{R}^m : \ ||x|| \leq 1\}$, and $\hat{\Sigma}_y r = L_y^T L_y$, for L_y upper triangular and $r^2 = mF_{m,n_i-1}(0.95)$. Then the projection of E onto the first two principal components of X is

$$P(E) = \{z | z = R(\bar{y} + L_y^{-T}u) \text{ for } u \in B_m\}$$

Constructing B_m is computationally wasteful, and absolutely impractical for large m. For this reason we seek an alternative involving only B_2.

Define $RL_y^{-T} = U_y D_y V_y^T$, the singular value decomposition. Then for and $u \in B_m$, $RL_y^{-T}u = U_y D_y V_y^T u = U_y D_y \tilde{u}$, for $\tilde{u} \in B_2$, since V^T is orthonormal. Thus

$$P(E) = \{z | z = R\bar{y} + U_y D_y \tilde{u} \text{ for } \tilde{u} \in B_2\}.$$

Best fit models for Terrapene trees

The best fit substitution models and partitioning schemes for each locus as given by PartitionFinder (Lanfear et al. 2012) is shown in Table 1.

Table 1: The best fit substitution models and partitioning schemes for each locus as given by PartitionFinder (Lanfear et al. 2012).

Locus	Model	Length	Type
AnonTB29	HKY	522	Nuclear
AnonTB73	HKY	659	Nuclear
Cytb	TN93+I	1070	Mitochondrial
Gapd	HKY	401	Nuclear
HNFAL	K80	753	Nuclear
ODC	HKY	420	Nuclear
R35	HKY+I	913	Nuclear
RAG	HKY	686	Nuclear
TGF	HKY	863	Nuclear
VIM	HKY+G	676	Nuclear

References

- Pope, S.B. (2008). Algorithms for Ellipsoids. *Technical Report FDA-08-01.*
- Lanfear, R. et al. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* 29(6). 1695-1701.