The Time-Marginalized Coalescent Prior for Hierarchical Clustering – Supplementary Material

September 19, 2012

1 Proofs

Lemma 1. *A tree ψ_n has $T(\psi_n) = \frac{(n-1)!}{\prod_{i=1}^{n-1} m_i}$ possible orderings on its internal nodes, where m_i is the number of internal nodes in the subtree rooted at node i.*

Proof. This can be done with a recursion relation. Consider an unordered tree ψ_n with n_l leaves in the left subtree, n_r in the right, and define $m_i = n_i - 1$ to be the number of internal nodes of subtree i. Additionally, there are k_l possible orderings of the left subtree's internal nodes, and k_r possible orderings of the right subtree's internal nodes. Consider the case where $k_l = k_r = 1$, where we need to count the number of ways in which the two sequences of length m_r and m_l can be "interleaved". There are $(m_r + m_l)!$ possible orderings of the $m_r + m_l$ nodes if we have no constraints on the orderings of the right nodes with respect to each other, and no such constraints for the left nodes. Thus, the total number of ways to interleave the two sequences is $\frac{(m_r+m_l)!}{m_r! m_l!} = \binom{m_r+m_l}{m_r}$. See Figure 1. For $k_r, k_l > 1$, we simply multiply these in, giving the total number of orderings as:

$$\binom{m_r + m_l}{m_r} k_r k_l$$

Note we can apply the same logic to the right and left subtrees to determine k_r and k_l. Given the children of node i are l_i and r_i, the number of orderings k_i consistent with a particular subtree i is:

$$k_i = \binom{m_{r_i} + m_{l_i}}{m_{r_i}} k_{r_i} k_{l_i}$$

So we can write out the total number of orderings as a product over all internal nodes of the tree:

$$T(\psi_n) = \prod_{i=1}^{n-1} \binom{m_{r_i} + m_{l_i}}{m_{r_i}}$$

Furthermore, note that $m_i = m_{l_i} + m_{r_i} + 1$ (the number of internal nodes in a tree is the sum of those in the subtrees, plus the root), so the denominator of the binomial coefficients of a parent node will cancel the numerators of the coefficients of its children, however leaving a term proportional to $\frac{1}{m_i}$, and the numerator of the root will not be canceled. Thus we have:

$$T(\psi_n) = \frac{(n-1)!}{\prod_{i=1}^{n-1} m_i}$$

□

Theorem 1. *$p(\psi_n)$ defines an exchangeable and consistent prior over Ψ_n*

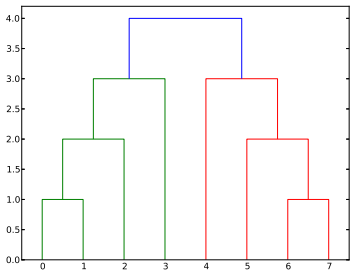

Figure 1: A tree with two unbalanced subtrees. The internal nodes for each of the subtrees can only be ordered in one way with respect to each other, however, for the overall tree there are $\frac{6!}{3!3!} = \binom{6}{3} = 20$ orderings of the internal nodes.

Proof. Since $p(\psi_n)$ doesn't depend on the order of data seen, it is exchangeable. Thus for consistency all we need to show is that we have a well defined conditional prior; ie that when we marginalize out a particular point from $p(\psi_n)$ we get back $p(\psi_{n-1})$, if ψ_{n-1} is a tree structure obtained by removing a leaf from ψ_n. We can do this by showing $\sum_{\psi_{n+1} \in C(\psi_n)} p(\psi_{n+1}) = p(\psi_n)$, where $C(\psi_n)$ is the set of ψ_{n+1} structurally consistent with a particular ψ_n. This is equivalent to showing:

$$\sum_{\psi_{n+1} \in C(\psi_n)} T(\psi_{n+1}) = \binom{n+1}{2} T(\psi_n)$$

which can be proven by induction, and we take this as our inductive hypothesis for all $l < n$. The base case for $n = 1$:

$$T(\psi_2) = \frac{(2-1)!}{1!1!} = 1 = \binom{2}{2} = \binom{2}{2} T(\psi_1)$$

Consider the tree ψ_n and its two subtrees ψ_{n_l} and ψ_{n_r}. As listed before, we have $T(\psi_n) = \binom{m_l+m_r}{m_l} T(\psi_{n_l}) T(\psi_{n_r})$. Depending on where we add the $n+1$st point, (denoted x_{n+1}) we have:

$$T(\psi_{n+1}) = \begin{cases} \binom{m_l+m_r+1}{m_l+1} T(\psi_{n_l+1}) T(\psi_{n_r}) & \text{if } x_{n+1} \text{ is added to the left subtree} \\ \binom{m_l+m_r+1}{m_r+1} T(\psi_{n_l}) T(\psi_{n_r+1}) & \text{if } x_{n+1} \text{ is added to the right subtree} \\ \binom{m_l+m_r+1}{m_l+m_r} \binom{m_l+m_r}{m_r} T(\psi_{n_l}) T(\psi_{n_r}) & \text{if } x_{n+1} \text{ is added above the root} \end{cases}$$

Summing over all such ψ_{n+1}, and invoking the inductive hypothesis, we have:

$$\begin{aligned} \sum_{\psi_{n+1} \in C(\psi_n)} T(\psi_{n+1}) = & \binom{m_l+m_r+1}{m_l+1} \binom{n_l+1}{2} T(\psi_{n_l}) T(\psi_{n_r}) \\ & + \binom{m_l+m_r+1}{m_r+1} \binom{n_r+1}{2} T(\psi_{n_l}) T(\psi_{n_r}) \\ & + (m_l+m_r+1) \binom{m_l+m_r}{m_l} T(\psi_{n_l}) T(\psi_{n_r}) \end{aligned}$$

$$
\begin{aligned}
&= \binom{m_l + m_r}{m_l} T(\psi_{n_l}) T(\psi_{n_r}) \left(\frac{m_l + m_r + 1}{m_l + 1} \frac{(n_l + 1) n_l}{2} + \frac{m_l + m_r + 1}{m_r + 1} \frac{(n_r + 1) n_r}{2} + 1 \right) \\
&= T(\psi_n) \left(\frac{(n-1)(n_l + n_r + 2)}{2} + 1 \right) \\
&= T(\psi_n) \left(\frac{(n-1)(n+2)}{2} + \frac{2}{2} \right) \\
&= \binom{n+1}{2} T(\psi_n)
\end{aligned}
$$

□

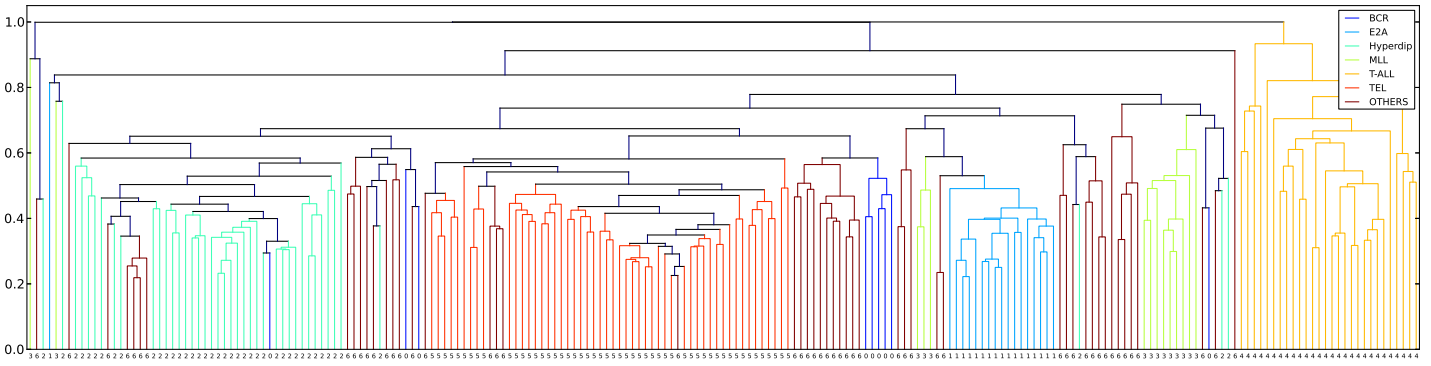

Figure 2: Posterior sample from our model applied to the leukemia dataset. Best viewed in color. Each pure subtree is painted a color unique to the class associated with it. The OTHERS class is a set of datapoints to which no diagnostic label was assigned.