



UvA-DARE (Digital Academic Repository)

The time-marginalized coalescent prior for hierarchical clustering

Boyles, L.; Welling, M.

Publication date

2013

Document Version

Submitted manuscript

Published in

26th Annual Conference on Neural Information Processing Systems 2012

[Link to publication](#)

Citation for published version (APA):

Boyles, L., & Welling, M. (2013). The time-marginalized coalescent prior for hierarchical clustering. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *26th Annual Conference on Neural Information Processing Systems 2012: December 3-6, 2012, Lake Tahoe, Nevada, USA* (Vol. 4, pp. 2969-2977). (Advances in Neural Information Processing Systems; Vol. 25). Curran Associates.
<https://papers.nips.cc/paper/4786-the-time-marginalized-coalescent-prior-for-hierarchical-clustering>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

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*

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

□

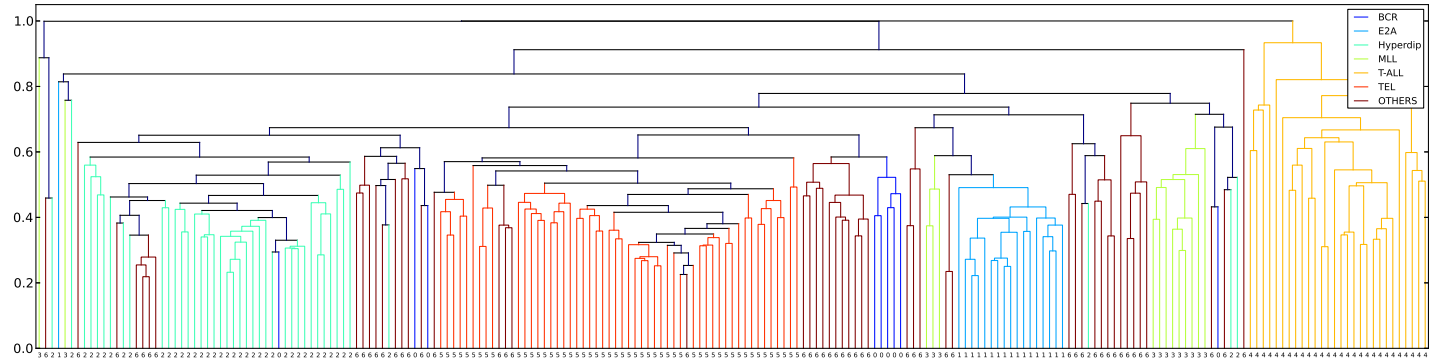


Figure 2: Posterior sample from our model applied to the leukemia dataset. Best viewed in color. Each pure subtree is painted a color unique to the class associated with it. The OTHERS class is a set of datapoints to which no diagnostic label was assigned.