Summary of "Bayesian Agglomerative Clustering with Coalescents"

Matt Dickenson `mcd31`

STA571/CS590.01 Due: 17 February, 2014

Teh, Daume and Roy (2008) introduce a new model for Bayesian agglomerative clustering based on a prior over trees derived from population genetics. The prior, known as Kingman's coalescent, describes genealogies in an evolutionary process. n haploid (single-parent) individuals observed at time $t = 0$ are assumed to have a common ancestor at $t = -\infty$. The genealogies of the n individuals form a directed forest, and $\pi(t)$ identifies the m ancestors of $[n] = \{1, \ldots, n\}$ at time t. A coalescent event refers to the point $t_i < 0$ when two individuals' genealogies converge at their common ancestor. The time between adjacent events δ_i is distributed exponentially with $\lambda = \binom{n-i+1}{2}$. Taking the limit as $n \to \infty$ we have a marginally independent, infinitely exchangeable distribution over genealogies known as the n-coalescent.

Inference for this model, using either sequential Monte Carlo or greedy algorithms, starts in the distant past and proceeds forward, splitting the ancestry of two data points later (closer to $t = 0$) the more closely dependent they are. The authors present several experiments comparing the coalescent model to related methods such as Bayesian Hierarchical Clustering (BHC). Coalescent inference outperforms BHC in most of their experiments, likely due to the fact that the coalescent shares information across the tree, whereas the mixture model underlying BHC is flat. BHC also does not define a distribution over trees as the n-coalescent does. The hierarchical component of this model also differs from the HDP insofar as the latter method uses a fragmentation process for trees, which is the reverse of the coalescent process. This makes inference for the n-coalescent simpler than existing HDP algorithms. A disadvantage of the coalescent method is that it is less efficient to compute than BHC ($O(n^2)$ for a greedy coalescent implementation versus $O(n \log n)$ for BHC). If computational efficiency is crucial, BHC would be preferable to the n-coalescent.