—PREPRINT—

Bayesian Agglomerative Clustering with Coalescents —Supplemental Material—

Yee Whye Teh
Gatsby Unit
University College London
ywteh@gatsby.ucl.ac.uk

Hal Daumé III
School of Computing
University of Utah
me@hal3.name

Daniel Roy
CSAIL
MIT
droy@mit.edu

Hyperparameter Estimation

In the Brownian motion case, the only hyperparameter in this model is the covariance matrix Λ. For simplicity, we consider only the diagonal case: $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_D)$. We place independent Gamma priors on the inverse variances with hyperparameters a and b. In our experiments we set $a = b = 1.1$ so that the prior has mode at 1. Conditioned on a geneology, the posterior distribution of λ_d^{-1} is again Gamma, with hyperparameters $\widehat{a}_d$ and $\widehat{b}_d$ given by:

$$\widehat{a}_d = a + \frac{1}{2}(n-1); \quad \widehat{b}_d^{-1} = b^{-1} + \frac{1}{2}\sum_{i=1}^{n-1} \frac{(\widehat{y}_{\rho_{li},d} - \widehat{y}_{\rho_{ri},d})^2}{v_{\rho_{li}} + v_{\rho_{ri}} + t_{li} + t_{ri} - 2t_i} \tag{1}$$

The MAP $\lambda_d^{-1} = (\widehat{a}_d - 1)/\widehat{b}_d$.

Next consider the binary vector case. The two hyperparameters q_{d1} and λ_d can be optimized separately for each entry d. Unfortunately there is no closed form solution and we used Newton steps, reparametrizing q_{d1} as $q_{d1} = 1/(1 + \exp(-v_d))$ so that the resulting optimization is unconstrained. The cost function to be maximized is:

$$\mathcal{L}_d(v_d, \lambda_d) = \sum_{i=1}^{n-1} \log\left(1 - e^{\lambda_h(2t_i - t_{li} - t_{ri})}\left(1 - (1-q_{d1})M^{d0}_{\rho_{li}}M^{d0}_{\rho_{ri}} - q_{d1}M^{d1}_{\rho_{li}}M^{d1}_{\rho_{ri}}\right)\right) \tag{2}$$

Updates for the multinomial case can be derived analogously.

Predictive Density

Given a tree and a new individual y' we wish to know: (a) where y' might coalescent and (b) what the density is at y'. To answer (a), assume that y' coalesces with the genealogy at time t, where $t_j > t > t_{j+1}$. The prior probability of this coalescesce is:

$$\exp[-\sum_{i=1}^{j}(n-i+1)\delta_i - (n-j)(t_j - t)] \tag{3}$$

At time t, there are $n-j$ individuals that y' could coalesce with. In the Brownian motion case, y' may merge with sibling ρ_s, and the parent of ρ_s is ρ_p. To perform this merge, we need to create a new parent $\rho_{p'}$ between ρ_s and ρ_p to become the parent of y' and ρ_s. The probability of this change is the probability of $\rho_{p'}$ under ρ_p, times the probability of ρ_s and y' under $\rho_{p'}$, divided by the probability of ρ_s under ρ_p. Marginalizing out $\rho_{p'}$, we obtain:

$$\left[(2\pi(v_0 - t))^D \det\Lambda\right]^{-1/2} \exp\left[-\frac{1}{2}\,||y_0 - y'||_{\Lambda(v_0-t)} - (n-j+1)(t_s - t)\right] \tag{4}$$

$$v_0 = [(v_{\rho_s} + t_s - t)^{-1} + (v_{\rho_p} + t - t_p)^{-1}]^{-1}; \quad y_0 = v_0[\hat{y}_{\rho_s}/(v_{\rho_s} + t_s - t) + \hat{y}_{\rho_p}/(v_{\rho_p} + t_p - t)]$$

Here, v_0 is the posterior variance and y_0 is the posterior mean; $\hat{y}_{\rho_s}$ and v_{ρ_s} are the messages passed *up* through the tree, while $\hat{y}_{\rho_p}$ and v_{ρ_p} are the messages passed *down* through the tree. The full predictive density is obtained by summing the product of the prior and Eq (4) over all siblings at all time steps; we draw 10 equally spaced samples between t_j and t_{j+1}. Care must be made to correctly handle the root: we draw 10 equally spaced samples beginning at the minimum t and $t - \max_j \delta_j$; moreover, there are no messages coming down from the root, so those terms are excluded from the likelihood in (4).

Marginal Likelihood Estimation

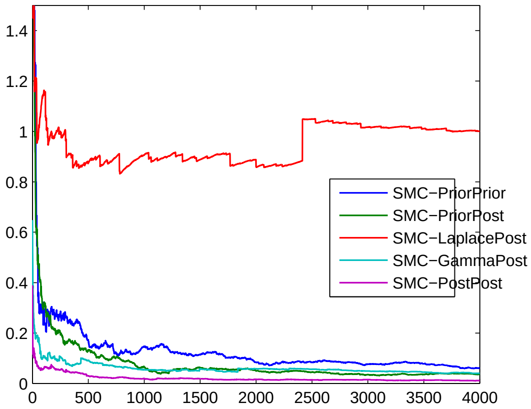

Figure 1: Error in Monte Carlo estimates of the marginal likelihood of a small data set.

In order to evaluate the quality of the proposal distributions, we calculated the exact marginal likelihood under the Brownian diffusion coalescent process on a small tree with data points at $\{-3.1416, 2.1718, 1.618\}$. We then ran the particle filters without resampling to gather 4000 weighted samples, computed the Monte Carlo estimate of the marginal likelihood for $n = 1, \ldots, 4000$, and measured the difference from the true marginal likelihood. Figure 1 shows the results. In summary, as expected, SMC-PostPost is the best. Instead of sampling from the coalescent time prior, and as an alternative to sampling from the computational expensive mixture of generalized inverse Gaussian in SMC-PostPost, various approximations to the conditional distribution on coalescent times were developed. A gaussian fit failed in this task, suffering from high variance. The gamma fit was superior, but in experience, also suffered from large variance. We believe both of these failed due to tails that are too short (the Gamma assigning too little mass close to zero).

NIPS Coalescent

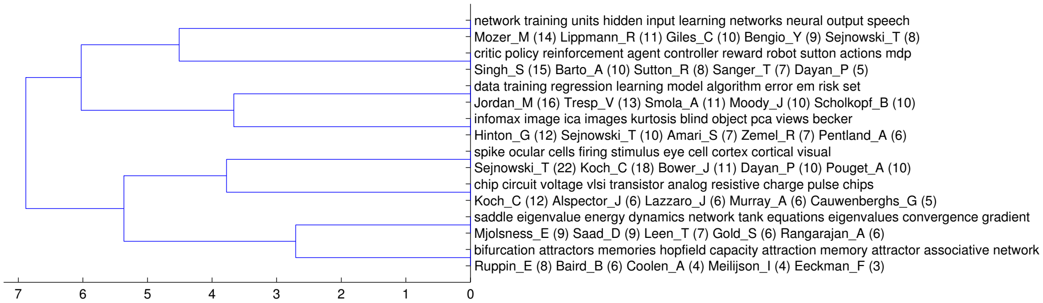
network training units hidden input learning networks neural output speech
Mozer_M (14) Lippmann_R (11) Giles_C (10) Bengio_Y (9) Sejnowski_T (8)
critic policy reinforcement agent controller reward robot sutton actions mdp
Singh_S (15) Barto_A (10) Sutton_R (8) Sanger_T (7) Dayan_P (5)
data training regression learning model algorithm error em risk set
Jordan_M (16) Tresp_V (13) Smola_A (11) Moody_J (10) Scholkopf_B (10)
infomax image ica images kurtosis blind object pca views becker
Hinton_G (12) Sejnowski_T (10) Amari_S (7) Zemel_R (7) Pentland_A (6)
spike ocular cells firing stimulus eye cell cortex cortical visual
Sejnowski_T (22) Koch_C (18) Bower_J (11) Dayan_P (10) Pouget_A (10)
chip circuit voltage vlsi transistor analog resistive charge pulse chips
Koch_C (12) Alspector_J (6) Lazzaro_J (6) Murray_A (6) Cauwenberghs_G (5)
saddle eigenvalue energy dynamics network tank equations eigenvalues convergence gradient
Mjolsness_E (9) Saad_D (9) Leen_T (7) Gold_S (6) Rangarajan_A (6)
bifurcation attractors memories hopfield capacity attraction memory attractor associative network
Ruppin_E (8) Baird_B (6) Coolen_A (4) Meilijson_I (4) Eeckman_F (3)
7
6
5
4
3
2
1
0

Figure 2: Top of the tree derived from the NIPS abstract data, with most indicative words and most frequent authors for each sub-node.

LLR ***(Time)***	**Top Words and** ***Top Authors***
32.7 *(-2.71)*	bifurcation attractors hopfield network saddle dynamics attractor eigenvalue equilibrium *Mjolsness_E (9) Saad_D (9) Ruppin_E (8) Coolen_A (7) Leen_T (7)*
.106 *(-3.77)*	voltage model cells neurons neuron cell figure spike input time *Koch_C (30) Sejnowski_T (22) Bower_J (11) Dayan_P (10) Pouget_A (10)*
83.8 *(-2.02)*	chip circuit voltage vlsi transistor analog resistive charge pulse chips *Koch_C (12) Alspector_J (6) Lazzaro_J (6) Murray_A (6) Cauwenberghs_G (5)*
140 *(-2.43)*	spike ocular cells firing stimulus eye cell cortex cortical visual *Sejnowski_T (22) Koch_C (18) Bower_J (11) Dayan_P (10) Pouget_A (10)*
2.48 *(-3.66)*	data model learning algorithm training set function latent mixture bayesian *Jordan_M (17) Hinton_G (16) Williams_C (14) Tresp_V (13) Moody_J (12)*
31.3 *(-2.76)*	infomax image ica images kurtosis blind object pca views becker *Hinton_G (12) Sejnowski_T (10) Amari_S (7) Zemel_R (7) Pentland_A (6)*
31.6 *(-2.83)*	data training regression learning model algorithm error em risk set *Jordan_M (16) Tresp_V (13) Smola_A (11) Moody_J (10) Scholkopf_B (10)*
39.5 *(-2.46)*	critic policy reinforcement agent controller reward robot sutton actions mdp *Singh_S (15) Barto_A (10) Sutton_R (8) Sanger_T (7) Dayan_P (5)*
23.0 *(-3.03)*	network training units hidden input learning networks neural output speech *Mozer_M (14) Lippmann_R (11) Giles_C (10) Bengio_Y (9) Sejnowski_T (8)*

Table 1: Nine clusters discovered in NIPS abstracts data.

Phylolinguistics

In the second experiment, we restrict ourselves to languages from the following families: Niger-Congo, Indo-European, Austronesian, Australian, Afro-Asiatic and Sino-Tibetan. We further require that a language have at most 60 of the 139 features unknown—this leaves 64 languages. The coalescent for these languages is shown—together with corresponding language families—in Figure 3. In this figure, we can see that the coalescent is able to identify almost all of Indo-European (with two exceptions: Persian is a bit far away and Hindi/Armenian are also). It does quite well with Austronesian languages, erring only with Paamese. The Australian languages are mixed up a bit with the Sino-Tibetan languages, which can perhaps be accounted for on the basis of areal sharing (i.e., language change due to close proximity).

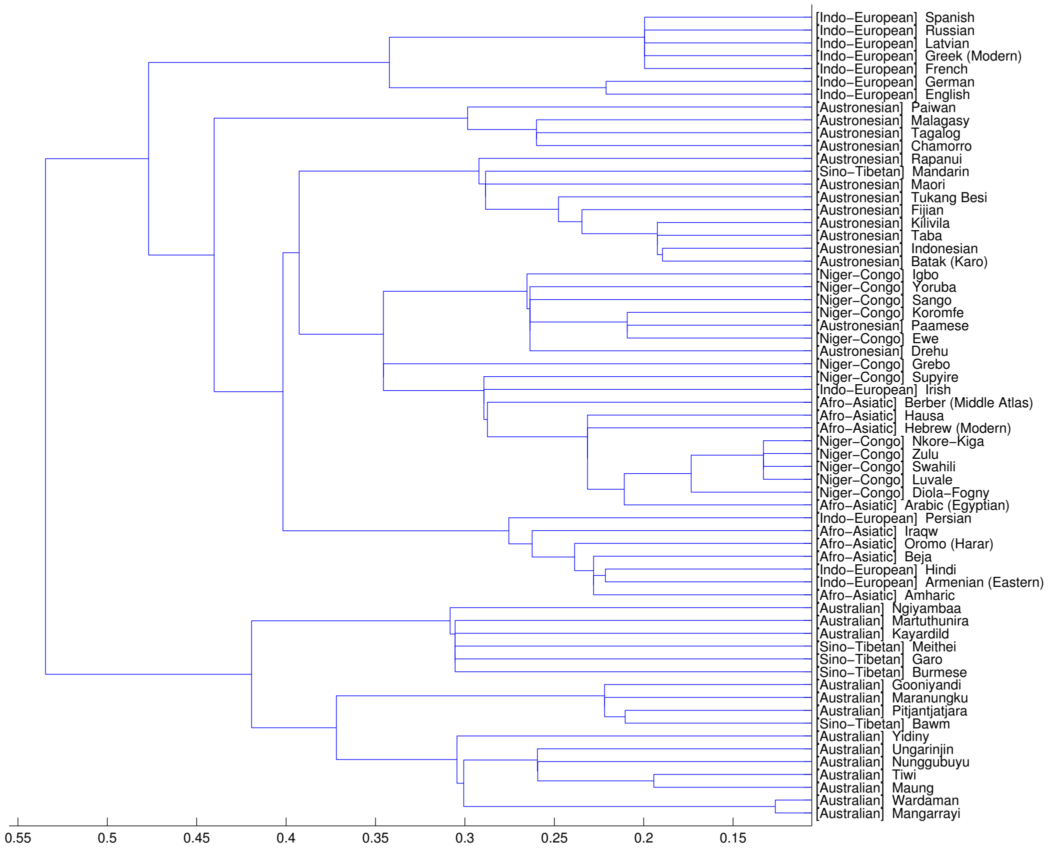

Figure 3: Coalescent for a subset of 64 languages from WALS.