The time-marginalized coalescent prior for hierarchical clustering

Boyles, L.; Welling, M.

Publication date
2013

Document Version
Final published version

Published in
26th Annual Conference on Neural Information Processing Systems 2012

Citation for published version (APA):

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.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date: 02 Nov 2021
Abstract

We introduce a new prior for use in Nonparametric Bayesian Hierarchical Clustering. The prior is constructed by marginalizing out the time information of Kingman’s coalescent, providing a prior over tree structures which we call the Time-Marginalized Coalescent (TMC). This allows for models which factorize the tree structure and times, providing two benefits: more flexible priors may be constructed and more efficient Gibbs type inference can be used. We demonstrate this on an example model for density estimation and show the TMC achieves competitive experimental results.

1 Introduction

Hierarchical clustering models aim to fit hierarchies to data, and enjoy the property that clusterings of varying size can be obtained by “pruning” the tree at particular levels. In contrast, standard clustering models must specify the number of clusters beforehand, while Nonparametric Bayesian (NPB) clustering models such as the Dirichlet Process Mixture (DPM) [5, 13] directly infer the (effective) number of clusters. Hierarchical clustering is often used in population genetics for inferring ancestral history and bioinformatics for genetic clustering, and has also seen use in computer vision [18, 1] and topic modelling [3, 1].

NPB models are a class of models of growing popularity. Being Bayesian, these models can easily quantify the uncertainty the resulting inferences, and being nonparametric, they can seamlessly adapt to increasingly complicated data, avoiding the model selection problem. NPB hierarchical clustering models are an important regime of such models, and have been shown to have superior performance to alternative models in many domains [8]. Thus, further advances in the applicability of these models is important.

There has been substantial work on NPB models for hierarchical clustering. Dirichlet Diffusion Trees (DDT) [16], Kingman’s Coalescent [9, 10, 4, 20], and Pitman-Yor Diffusion Trees (PYDT) [11] all provide models in which data is generated from a Continuous-Time Markov Chain (CTMC) that lives on a tree that splits (or coalesces) according to some continuous-time process. The nested CRP and DP [3, 17] and Tree-Structured Stick Breaking (TSSB) [1] define priors over tree structures from which data is directly generated.

Although there is extensive and impressive literature on the subject demonstrating its useful clustering properties, NPB hierarchical clustering has yet to see widespread use. The expensive computational cost typically associated with these models is a likely inhibitor to the adoption of these models. The CTMC based models are typically more computationally intensive than the direct generation models, and there has been substantial work in improving the speed of inference in these models. [12] introduces a variational approximation for the DDT, and [7, 6] provide more efficient SMC schemes for the Coalescent. The direct generation models are typically faster, but usually
A pair is uniformly drawn from \(N = 5\) points to coalesce. The coalescence time \(t\) is drawn from \(\text{Exp}(\frac{1}{2})\), and another pair on the remaining 4 points is drawn uniformly. After drawing \(t_4 \sim \text{Exp}(\frac{1}{2})\), the coalescence time for the newly coalesced pair is \(t_5 + t_4\).
necessary because for a given $\psi_n$, there are multiple ways of constructing it by uniformly picking pairs to join, see Figure 2. If there are $i$ remaining nodes to join, there are $\binom{i}{2}$ ways of joining them, so we have for the probability of a particular $\phi_n$:

$$p(\phi_n) = \prod_{i=2}^{n} \left(\binom{i}{2}\right)^{-1}$$

This defines an exchangeable and consistent prior over $\Phi_n$: exchangeable because $p(\phi_n)$ does not depend on the order in which the data is seen, and consistent because the conditional prior\footnote{The sequential sampling scheme often associated with NPB models; for example the conditional prior for the CRP is the prior probability of adding the $n+1$st point to one of the existing clusters (or a new cluster) given the clustering on the first $n$ points.} is well defined – we can imagine adding a new leaf to an existing $\phi_n$, which creates a new internal node. Let $y_i$ denote the $i$th internal node\footnote{When times are given, we index the internal nodes from most recent to root. Otherwise, nodes are ordered such that parents always succeed children.} of $\phi_n$, $i \in \{1...n-1\}$, and let $y^*$ denote the new internal node. There are $n$ ways of attaching the new internal node below $y_1$, $n-1$ ways of attaching below $y_2$, and so on, giving $\frac{n(n+1)}{2}$ ways of attaching $y^*$ into $\phi_n$. Thus if we make this choice uniformly at random, we get the probability of the new tree is

$$p(\phi_{n+1}) = p(\phi_n)\left(\frac{n+1}{2}\right)^{-1} = \prod_{i=2}^{n+1} \left(\binom{i}{2}\right)^{-1}$$

It is possible to marginalize out the ordering information in the coalescent tree structures $\phi_n$ to derive exchangeable, consistent priors on “unordered” tree structures $\psi_n$. We can perform this marginalization by counting how many ordered tree structures $\phi_n \in \Phi_n$ are consistent with a particular unordered tree structure $\psi_n$.

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

(For proof see the supplementary material.) This is in agreement with what we would expect: for an unbalanced tree, $m_i = \{1, 2, ..., n-1\}$, so this gives $T = 1$. Since an unbalanced tree imposes a full ordering on the internal nodes, there can only be one unbalanced ordered tree that maps to the corresponding unbalanced unordered tree. As the tree becomes more balanced, the $m_i$s decrease, increasing $T$.

Thus the probability of a particular $\psi_n$ is $T(\psi_n)$ times the probability of an ordered tree $\phi_n$ under the coalescent:\footnote{It has been brought to our attention that this prior and its connection to the coalescent has been studied before in [2] as the beta-splitting model with parameter $\beta = 0$, and later in [14] under the framework of Gibbs-Fragmentation trees.}

$$p(\psi_n) = T(\psi_n) \prod_{i=2}^{n} \left(\binom{i}{2}\right)^{-1} = \frac{(n-1)!}{\prod_{i=1}^{m_i}} \prod_{i=2}^{n} \left(\binom{i}{2}\right)^{-1}$$

\[(1)\]

**Theorem 1.** $p(\psi_n)$ defines an exchangeable and consistent prior over $\Psi_n$. $p(\psi_n)$ is clearly still exchangeable as it does not depend on any order of the data, and was defined by marginalizing out a consistent process, so its conditional priors are still well defined and thus $p(\psi_n)$ is consistent. For a more explicit proof see the supplementary material.

Figure 3: (left) A sample from the described prior with stick-breaking parameter $B(1, 1)$ (uniform). (middle) A sample using $B(2, 2)$. (right) A sample using $B(4, 2)$. 
Figure 4: The subtree $S_l$ rooted at the red node $l$ is pruned in preparation for an MCMC move. We perform slice sampling to determine where the pruned subtree’s parent should be placed next in the remaining tree $\pi_l$.

Figure 5: We compute the posterior pdf for each branch that we might attach to. If $\alpha$ or $\beta$ are greater than one, the Beta prior on branch lengths can cause these pdfs to go to zero at the limits of their domain. Thus to enable moves across branches we compute the extrema of each pdf so that all valid intervals are found.

3 Data Generation Model

Given a prior over tree structures such as (1), we can define a data generating process in many ways (indeed, any $L_1$ bounded martingale will do [19]); here we restrict our attention to generative models in which we first sample times given a tree structure, and then sample the data according to some process described on those times (in our case Brownian Motion). Examples of other potential data generation models include those in [1], such as the “Generalized Gaussian Diffusions,” and the multinomial likelihood often used with the Coalescent.

3.1 Branch Lengths

Given a tree structure $\psi$, we can sample branch lengths, $s_i = t_{p_i} - t_i$, with $t_i$ the time of coalescent event $i$, with $t = 0$ at the leaves and $p_i$ is the parent of $i$. Consider the following construction similar to a stick-breaking process: Start with a stick of unit length. Starting at the root, travel down the given $\psi$, and at each split point duplicate the current stick into two sticks, assigning one to each child. Then, sample a Beta random variable $B$ for each of the two sticks where the corresponding children are not leaves. $B$ will be the proportion of the remaining stick attributed to that branch of the tree until the next split point (sticks afterwards will be of length proportional to $(1 - B)$). We have, $B_i = 1 - \left(\frac{t_i}{t_{p_i}}\right)$ = $\frac{s_i}{t_{p_i}}$. The total prior over branch lengths can thus be written as:

$$p(\{B_i\} | \psi) = \prod_{i=1}^{N-2} B(B_i | \alpha, \beta)$$ (2)

See Figure 3 for samples from this prior. Note that any distribution whose support is the unit interval may be used, and in fact more innovative schemes for sampling the times may be used as well; one of the major advantages of the TMC over the Coalescent and DDT is that the times may be defined in a variety of ways.

There is a single Beta random variable attributed to each internal node of the tree (except the root, which has $B$ set to 1). Since the order in which we see the data does not affect the way in which we sample these stick proportions, the process remains exchangeable. We denote pairs $(\psi_n, \{B_i\})$ as $\pi$, i.e. a tree structure with branch lengths.

3.2 Brownian Motion

Given a tree $\pi$ we can define a likelihood for continuous data $x_i \in \mathbb{R}^p$ using Brownian motion. We denote the length of each branch segment of the tree $s_i$. Data is generated as follows: we start at some unknown location in $\mathbb{R}^p$ at time $t = 1$ and immediately split into two independent Wiener processes (with parameter $\Lambda$), each denoted $y_i$, where $i$ is the index of the relevant branch in $\pi$. Each
of the processes evolves for times $s_i$, then, a new independent Wiener process is instantiated at the
time of each split, and this continues until all processes reach the leaves of $\pi$ (i.e. $t = 0$), at which
point the $y_i$s at the leaves are associated with the data $x$. This is a similar likelihood to the ones used

### 3.2.1 Likelihood Computation

The likelihood $p(x|\pi)$ can be calculated using a single bottom-up sweep of message passing. As in
[20], by marginalizing out from the leaves towards the root, the message at an internal node $i$ is a
Gaussian with mean $\hat{y}_i$ and variance $\Lambda \hat{y}_i$.

The $\nu$ and $\hat{y}$ messages can be written for any number of incoming nodes in a single form:

$$\nu_i^{-1} = \sum_{j \in c(i)} (\nu_j + s_j)^{-1}; \quad \hat{y}_i = \nu_i \sum_{j \in c(i)} \frac{\hat{y}_j}{\nu_j + s_j}$$

where $c(i)$ are the nodes sending incoming messages to $i$. We can compute the likelihood using any
arbitrary node as the root for message passing. Fixing a particular node as root, we can write the
total likelihood of the tree as:

$$p(x|\pi) = \prod_{i=1}^{n-1} Z_{c(i)}(x, \pi)$$

When $|c(i)| = 1$ (e.g. when passing through the root at $t = 1$), $Z_{c(i)} = 1$. When $|c(i)| = 2$ and
$|c(i)| = 3$ (when collecting at an arbitrary node $i$ chosen as the root):

$$Z_{l_{ii_{i}}, r_{i}}(x, \pi) = [2\pi \hat{\Lambda}_i]^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \frac{1}{\hat{\Lambda}_i} [\hat{y}_r - \hat{y}_l]^2 \right); \quad \hat{\Lambda}_i = \hat{\Lambda} + \nu_i + s_i$$

$$Z_{P_{i}, l_{i}, r_{i}}(x, \pi) = [2\pi \Lambda]^{-1} \nu^* \frac{1}{2} \nu^* e^{\frac{1}{2} \nu^* \frac{1}{2} (\nu^* \frac{1}{2} \nu^* + \nu^* \frac{1}{2} \nu^*)}$$

$$\nu^* = \nu_{p_i} + s_i; \quad \nu^*_i = \nu_{l_i} + s_i; \quad \nu^*_r = \nu_{r_i} + s_i; \quad \nu^* = \nu^*_{p_i} + \nu^*_{l_i} + \nu^*_r + \nu^*_{p_i}$$

where $||.||_\Lambda$ corresponds to the Mahalanobis norm with covariance $\Lambda$. These messages are derived
by using the product of Gaussian pdf identities.
3.3 MCMC Inference

We propose an MCMC procedure that samples from the posterior distribution over $\pi$ as follows. First, a random node $l$ is pruned from the tree (so that its parent $p_l$ has no parent and only one child), giving the pruned subtree $S_l$ and remaining tree $\pi_l$. See Figure 4. We then consider all possible moves that would place $p_l$ into a valid location elsewhere in the tree. For each branch indexed by the node $i$ “below” it, we compute the posterior density function of where $p_l$ should be placed on that branch. We then slice sample on this collection of density functions. See Figure 5. By cycling through the nodes to prune and reattach, we achieve a Gibbs sampler over $\pi$.

We can efficiently compute the relative change in the likelihood $p(x|\pi)$ through a combination of belief propagation and local computations. First we perform belief propagation on $\pi_l$ to give upward and downward messages, and on $S_l$ to give only upwards messages. Denote $\pi(S, i, t)$ as the tree formed by attaching $S$ above node $i$ at time $t$ in $\pi$. For the new tree we imagine collecting messages to node $p_i$ resulting in a new factor $Z_{i, p_i}(x, \pi_l(S_l, i, t))$. The messages directly downstream of this factor are $Z_{\alpha(i)}(x, \pi_l)$ and $Z_{p_i, r_i}(x, \pi_l)$ (if $l_{p_i} = i$, i.e., $i$ is the “left” child of its parent). If we now imagine that the original likelihood was computed by collecting to node $p_i$, then we see that the first factor should replace the factor $Z_{p_i, r_i}(x, \pi_l)$ at node $p_i$ while the latter factor was already included in $Z_{l_{p_i}, r_{p_i}, p_i}(x, \pi_l)$. All other factors do no change. The total (multiplicative) change in the likelihood is thus,

$$\Delta Z(\pi_l(S_l, i, t)) = \frac{Z_{i, p_i}(x, \pi_l(S_l, i, t))}{Z_{l_{p_i}, r_{p_i}, p_i}(x, \pi_l)} Z_{p_i, r_i}(x, \pi_l)$$  \hspace{1cm} (6)

The update in prior probability for adding the parent of $l$ in the segment $(i, p_i)$ (with times $t_i$ and $t_{p_i}$) at time $t$ is proportional to the overall prior probability over $\psi_{\alpha, \beta}$

$$p(\pi_l(S_l, i, t)) \propto \frac{1}{B(1 - \frac{t}{t_{p_i}}; \alpha, \beta)} B(1 - \frac{t_i}{t}; \alpha, \beta) B(1 - \frac{t_i}{t}; \alpha, \beta) p(\psi(\pi_l(S_l, i, t)))$$  \hspace{1cm} (7)

Where $\psi(\pi)$ gives the structure part of $\pi = (\psi, \{B_i\})$. $p(\psi(\pi_l(S_l, i, t)))$ can be computed for all $i$ in linear time via dynamic programming (it does not depend on the actual value of $t$). By taking the product of (6) and (7) we get the joint posterior of $(i, t)$:

$$p(\pi_l(S_l, i, t)|X) \propto \Delta Z(\pi_l(S_l, i, t)) p(\pi_l(S_l, i, t))$$  \hspace{1cm} (8)

$p(\pi_l(S_l, i, t)|X)$ defines the distribution from which we would like to sample. We propose a slice sampler that can propose adding $S_l$ to any segment in $\pi_l$. For a fixed $i$, $p(\pi_l(S_l, i, t)|X)$ is typically unimodal, and typically has a small number of modes at most. If we can find all of the extrema of the posterior, we can easily find the intervals that contain positive probability density for slice sampling moves (see Figure 5). Thus this slice sampling procedure will mix as quickly as slice sampling on a single unimodal distribution. We find the extrema of these functions using Newton methods.

The overall sampling procedure is then to sample a new location for each node (both leaves and internal nodes) of the tree using the Gibbs sampling scheme explained above.

3.4 Hyperparameter Inference

As we do not know the structure of the data beforehand, we may not want to predetermine the specific values of $\alpha, \beta$ and $\Lambda$. Thus we define hyperpriors on these parameters and infer them as well. For simplicity we assume the form $\Lambda = kI$ for the Brownian motion covariance parameter. We use an Inverse-Gamma prior on $k$, so that $k^{-1} \sim \mathcal{G}(\kappa, \theta)$.

$$k^{-1}|X \sim \mathcal{G} \left( \frac{(N - 1)p}{2} + \kappa, \frac{1}{2} \sum_{i=1}^{N-1} d_i + \theta \right)$$

\[ \text{Note that if either } l \text{ or } i \text{ is a leaf, then the prior term will be simpler than the one listed here.} \]

6
where \( N \) is the number of datapoints, \( p \) is the dimension, and 
\[
d_i = \frac{||\hat{y}_i - \hat{y}_{r_i}||^2}{\nu_i + \nu_{r_i} + s_i + s_{r_i}},
\]
\( ||.|| \) is the Euclidean norm.

By putting a \( G(\xi, \lambda) \) prior on \( \alpha - 1 \) and \( \beta - 1 \), we achieve a posterior for these parameters:
\[
p(\alpha, \beta|\xi, \lambda, X) \propto (\alpha - 1)^{\xi}(\beta - 1)^{\xi} e^{-\lambda(\alpha - 1 + \beta - 1)} \prod_{i=1}^{N-1} \frac{1}{B(\alpha, \beta)} \left(1 - \frac{t_i}{\lambda p_i}\right)^{\alpha - 1} \left(\frac{t_i}{\lambda p_i}\right)^{\beta - 1}
\]

This posterior is log-concave and thus unimodal. We perform slice sampling to update \( \alpha \) and \( \beta \).

### 3.5 Predictive Density

Given a set of samples from the posterior, we can approximate the posterior density estimate by sampling a test point located at \( y_t \) into each of these trees repeatedly (giving new trees \( \pi' = \pi(y_t, i, s) \) for various values of \( i \) and \( s \)) and approximating \( p(y_t|X) \) as:

\[
p(y_t|X) = \int p(\pi|X)p(y_t|\pi)d\pi = \int d\pi p(\pi|X) \int d\pi' p(\pi', y_t|\pi, X)
\]

\[
\approx \sum_{\pi_1 \sim \pi|X} \sum_{\pi' \sim p(\pi|\pi_1, X)} p(y_t|\pi'_1, X)
\]

where \( p(\pi'|\pi, X) = \int p(\pi', y_t|\pi, X)dy_t \). By integrating out \( y_t \) in (5), we get a modification of (8) that is proportional to \( p(\pi'|\pi, X) \). Slice sampling from this gives us several new trees \( \pi'_1 \) for each \( \pi_i \), where one of the leaves is not observed. \( p(y_t|\pi'_1, X) \) is then available by message passing to the leaf associated with \( y_t \) (denoted \( l \)), which results in a Gaussian over \( y_t \). Thus the final approximated density is a mixture of Gaussians with a component for each of the \( \pi'_1 \).

Performing the aforementioned integration (after replacing \( \hat{y}_i \) with \( y_t \)), we get:

\[
Z_{pred}(i, t) \propto \int dy_t Z_{i, p_i, t}(x, \pi'(y_t))
\]

\[
= |2\pi\Lambda|^{-\frac{1}{2}}(\nu^*_i + \nu^*_{p_i})^{-\frac{1}{2}} \exp \left(-\frac{1}{2}(\nu^*_i + (\nu^*_i)^{-1} + \nu^*_{p_i} - (\nu^*_{p_i})^{-1})d_i, p_i\right)
\]

where \( d_i, p_i = ||\hat{y}_i - \hat{y}_{p_i}||^2_{\Lambda^{\nu_{r_i}}} \). This gives the posterior density for the location of the unobserved point:

\[
p(\pi' = \pi\{i\}, i, t|\pi, X) \propto Z_{pred}(i, t) \frac{Z_{i, r_i, t}(x, \pi)}{Z_{i, r_i, p_i}(x, \pi')} p(\pi\{i\}, i, t)
\]

where \( p(\pi\{i\}, i, t) \) is as in (7).

### 4 Experiments

We compare our model to Dirichlet Diffusion Trees (DDT) [16] and to Dirichlet Process Mixtures (DPM) [5, 13]. We used Radford Neal’s Flexible Bayesian Modeling package [15] for both the DDT and the DPM experiments. All algorithms were run with vague hyperpriors, except for the DPM concentration parameter which we set to .1 as we did not expect many modes for these experiments.

#### 4.1 Synthetic Data

To qualitatively compare our method to Dirichlet Process Mixtures and Dirichlet Diffusion Trees, we ran all three methods on a simulated dataset with \( N = 200, p = 2 \). The data is generated from a mixture of heavy tailed distributions to demonstrate the differences between these algorithms when presented with outliers. As can be seen in Figure 6, the DDT fits a density with reasonably heavy tails, whereas our model fits a narrower distribution. This is a result of the fact that the divergence function of the DDT strongly encourages the branch lengths to be small, and thus a larger variance is required to explain the data. Our model can be combined with a heavy-tailed observation model to produce densities with heavier tails – see the rightmost panel of Figure 6.
4.2 Gene Clustering

We applied our model to the St. Jude’s Leukemia dataset [22], which has $N_{\text{train}} = 215$ datapoints, $N_{\text{test}} = 112$, and preprocessed\(^5\) to have $p = 1000$ dimensions. We preprocessed the data so that each dimension had unit variance. Associated with each datapoint is one of 6 classifications of leukemia, or a 7th class with which no diagnosis was attributed. We applied our method to the full dataset to see if it could recover these classes. Figure 7 shows the posterior tree sampled after about 28 Gibbs passes (about 10 minutes). We also compared our method against the DDT and DPM on these models’ abilities to predict test data on a $p = 200$ subset of the $p = 1000$ dataset, as well as on the $p = 1000$ dataset, see Figure 8. On the $p = 200$ dataset, both the DDT and the TMC outperform the DPM, with the TMC performing slightly worse. We attribute this difference in performance due our model’s weaker prior on the branch lengths, which causes our model to overfit slightly; if we preset the diffusion variance of our model to a value somewhat larger than the data variance, our performance improves. In the $p = 1000$ dataset, the same phenomenon is observed.

4.3 Computer Vision Features

We also cluster visual bag of words features collected from birds images from Visipedia [21]. We worked on a dataset of size $N = 1400$, $N_{\text{test}} = 1412$, where each observation belongs to one of 200 classes of birds, see Figure 8. Again our method is better than DPM yet not as well as the DDT. Fixing the variance does improve the performance of our algorithm but not enough to improve over the DDT.

5 Conclusion

We introduced a new prior for use in NPB hierarchical clustering, one that can be used in a variety of ways to define a generative model for data. By marginalizing out the time of the coalescent, we achieve a prior from which data can be either generated directly via a graphical model living on trees, or by a CTMC lying on a distribution over times for the branch lengths – in the style of the coalescent and DDT. However, unlike the coalescent and DDT, in our model the times are generated conditioned on the tree structure; giving potential for more interesting models or more efficient inference. The simplicity of the prior allows for efficient Gibbs style inference, and we provide an example model and demonstrate that it can achieve similar performance to that of the DDT. However, to achieve that performance the diffusion variance must be set in advance, suggesting that alternative distributions over the branch lengths may provide better performance than the one explored here.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 0914783, 0928427, 1018433, 1216045.

\(^5\)We simply took the 1000 dimensions with the highest variance.
References


