Properties of Bayesian nonparametric models and priors over trees

David A. Knowles
Computer Science Department
Stanford University

July 24, 2013

Introduction

Theory: what characteristics might we want?

- exchangeability
- projectivity/consistency
- large support

$\rightarrow$ de Finetti's representation theorem

Applications: distributions over trees

- The nested Chinese restaurant process and the Dirichlet diffusion tree
- Kingman's coalescent

If we have time: fragmentation-coagulation processes.

Exchangeability

- Intuition: Data = pattern + noise
- Consider data $X_1, X_2, ...$ where $X_i \in \mathcal{X}$.
- A very strong assumption we could make: X_i are i.i.d. (*in the Bayesian paradigm* this corresponds to data=noise)
- An often reasonable assumption: X_i are *exchangeable*

$X_1, X_2, ...$ are *exchangeable* if $P(X_1, X_2, ...)$ is invariant under any finite permutation σ, i.e.

$$P(X_1, X_2, ...) = P(X_{\sigma(1)}, X_{\sigma(2)}, ...)$$

We can still have dependence!
Intuition: the order of the observations doesn't matter.

Exchangeability: the CRP

Recall the CRP predictive probability:

$$P(c_n = k | c_1, ..., c_{n-1}) = \begin{cases} \frac{a_{k,n-1}}{n-1+\theta} & k \in \{1, \ldots, K_n\} \\ \frac{\theta}{n-1+\theta} & k = K_n + 1 \end{cases}$$

where K_n is the number of blocks after assigning c_n and $a_{k,n} = \sum_{i=1}^{n} \mathbb{I}[c_i = k]$. Easy to show that

$$\begin{aligned} P(c_1, ..., c_n) &= \frac{\theta^{K-1} \prod_{k=1}^{K} 1 \cdot 2 \cdots (a_{k,n} - 1)}{(\theta + 1) \cdots (\theta + n - 1)} \\ &= \frac{\Gamma(\theta)\theta^K}{\Gamma(\theta + n)} \prod_{k=1}^{K} \Gamma(a_{k,n}) \end{aligned}$$

using $\Gamma(a + 1) = a\Gamma(a)$. Just depends on the size and number of blocks, and n. No dependence on the order: exchangeable! (but dependence...)

Exchangeability: breaking the CRP

I want more reinforcement! Let's square those counts.

$$P(c_n = k | c_1, ..., c_{n-1}) = \begin{cases} \frac{a_{k,n-1}^2}{\sum_{i=1}^K a_{i,n-1}^2 + \theta} & k \in \{1, \ldots, K_n\} \\ \frac{\theta}{\sum_{i=1}^K a_{i,n-1}^2 + \theta} & k = K_n + 1 \end{cases}$$

But no longer exchangeable :(

$$P(c_1, ..., c_n) = \frac{\theta^{K-1} \prod_{k=1}^K 1^2 \cdot 2^2 \cdots (a_{k,n} - 1)^2}{(\theta + 1)(\theta + \sum_{i=1}^{K_2} a_{i,2}^2) \cdots (\theta + \sum_{i=1}^{K_{n-1}} a_{i,n-1}^2)}$$

Projectivity

- This is a *consistency* property.
- Consider the finite sequence of random variables $X_1, X_2, ..., X_N$, with law P_N.
- Projectivity requires that

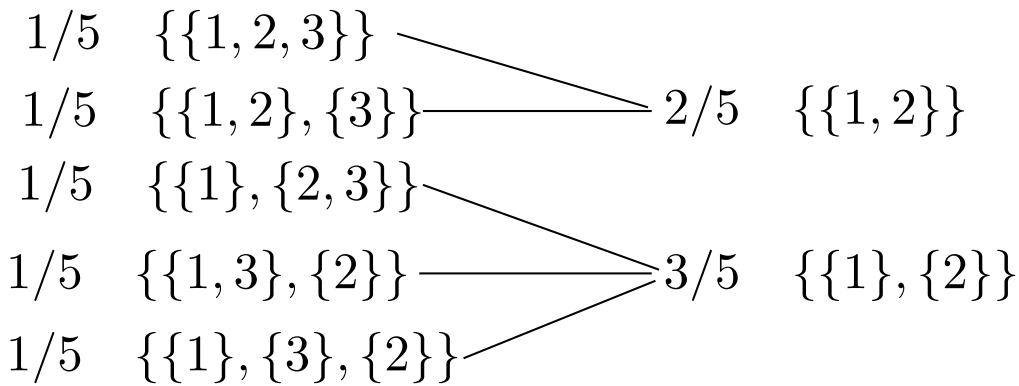
$$P_{N-1}(X_1, ..., X_{N-1}) = \int P_N(X_1, X_2, ..., X_N) dX_N$$

- Priors with a sequential generative process get this "for free" (e.g. the CRP):

$$P_N(X_1, X_2, ..., X_N) = P_{N-1}(X_1, ..., X_{N-1}) P(X_N | X_1, ..., X_{N-1})$$

Projectivity: a counterexample

The uniform distribution on partitions is exchangeable but *not projective*:



de Finetti's theorem

- Kolmogov extension theorem: if P_N are projective then there exists a limiting distribution P on the infinite sequence $X_1, X_2, ...$ whose finite dimensional marginals are given by P_N
- de Finetti's theorem: in that case, if $X_1, X_2, ...$ are exchangeable then there exists Q s.t.

$$P(X_1, X_2, ...) = \int_{M(\mathcal{X})} \prod_i \mu(X_i) Q(d\mu)$$

  where $M(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$ and $\mu \in M(\mathcal{X})$ is a random probability measure.
- As a graphical model:

$$\mu \sim Q$$
$$X_i|\mu \sim^{iid} \mu \qquad \forall i$$

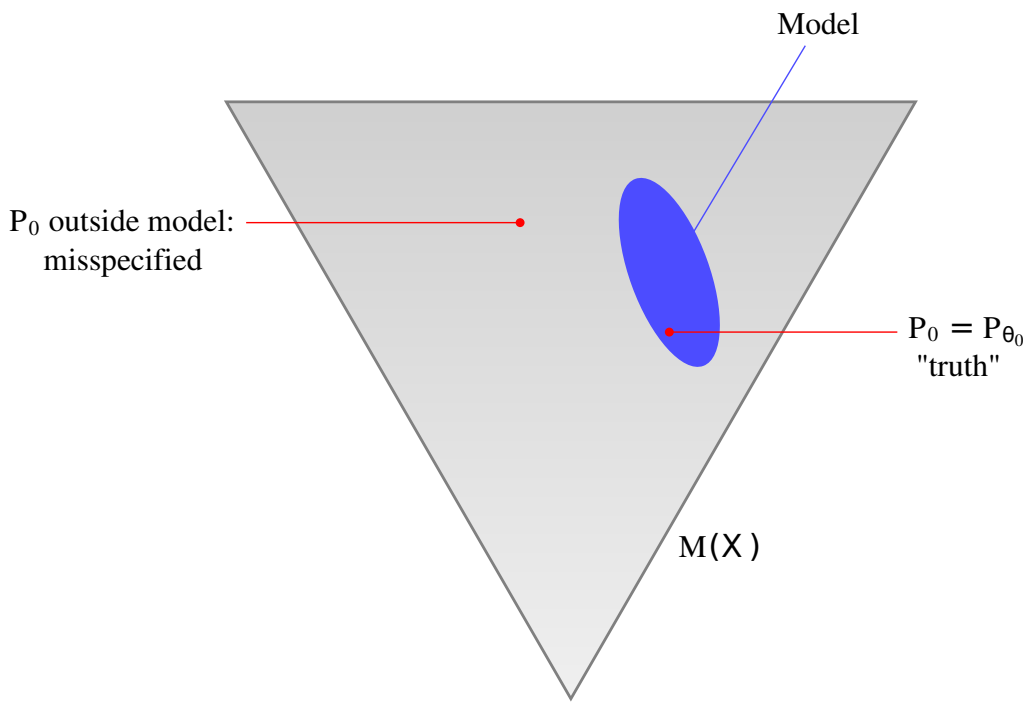
- **In words**: any exchangeable sequence of r.v.s can be represented as (formally, is equal in distribution to) a *mixture of i.i.d. r.v.s.*

de Finetti's theorem: notes

- Q is a "distribution over distributions": sometimes called the *de Finetti mixing measure*
- Not a constructive proof: sometimes we can characterise Q nicely (e.g. the Dirichlet process for the CRP, the beta process for the Indian buffet process), sometimes not (e.g. the tree priors presented later)
- Decomposition into pattern μ and noise
- Motivates Bayesian hierarchical models: Q is the prior
- μ is ∞-dimensional in general but might be finite in some cases (e.g. proportion ones for binary sequence)
- Just one example of a family of "ergodic decomposition theorems": group invariance $\rightarrow$ integral decomposition (e.g. Aldous-Hoover theorem for random graphs or arrays)
- Can get more formal: μ is determined by the limiting empirical distribution of the data, or the tail-σ-algebra of the sequence X_i

Support and consistency

- The space of probability measures $M(\mathcal{X})$ might be finite (e.g. for $\mathcal{X} = \{1, 2\}$) or infinite (e.g. for $\mathcal{X} = \mathbb{N}$ or $\mathbb{R}$)
- Nonparametric models can have support (non-zero probability mass) on infinite dimensional $M(\mathcal{X})$ whereas a parametric model can only put support on a finite-dimensional subspace [figure from Teh and Orbanz (2011)]



Consistency

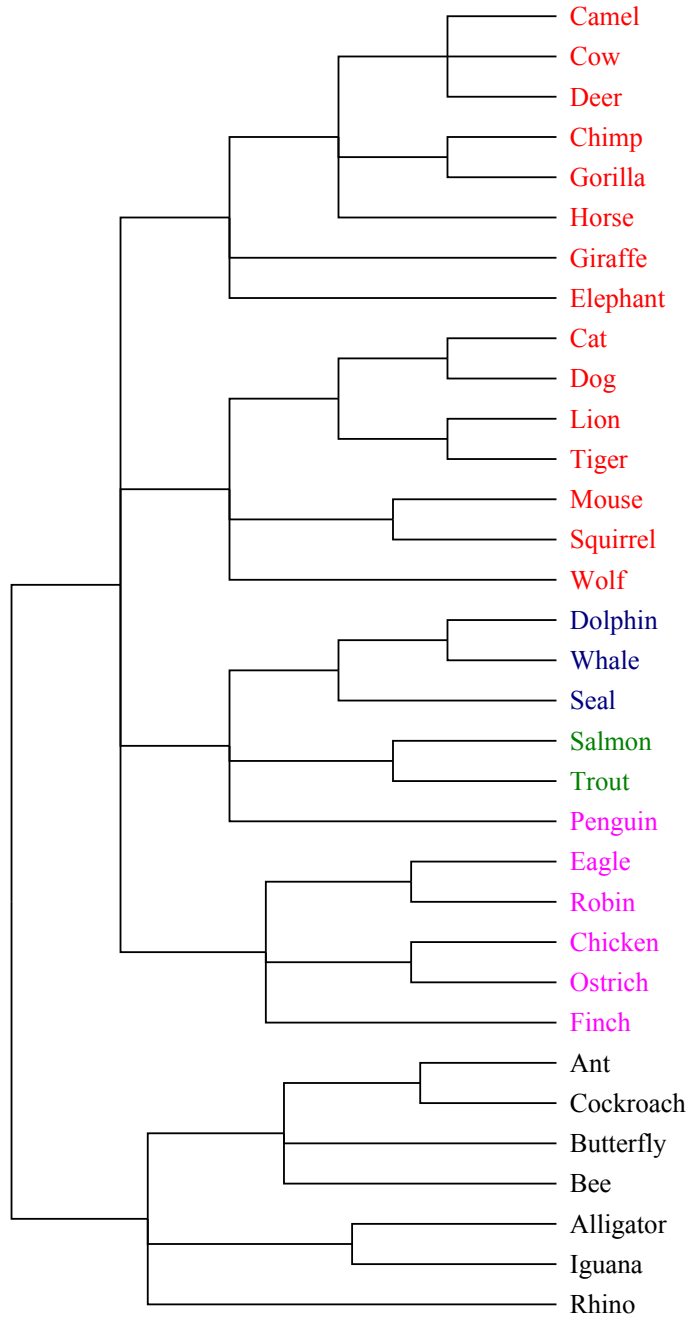
- Under mild conditions (e.g. identifiability) Bayesian models exhibit *weak* consistency: in the infinite data limit the posterior will converge to a point mass at the correct parameter value (μ), assuming this was sampled from the prior (says nothing about model misspecification, or approximate inference)
- Frequentist consistency (consistency for some true $P_0 \in M(\mathcal{X})$ in some class) is more difficult to ensure: some cases are known, e.g. Dirichlet process mixtures of diagonal covariance Gaussians for smooth densities.
- Convergence *rates* are also an active area of research: smooth parametric models typically get $n^{-\frac{1}{2}}$

Break

Next up: priors over trees.

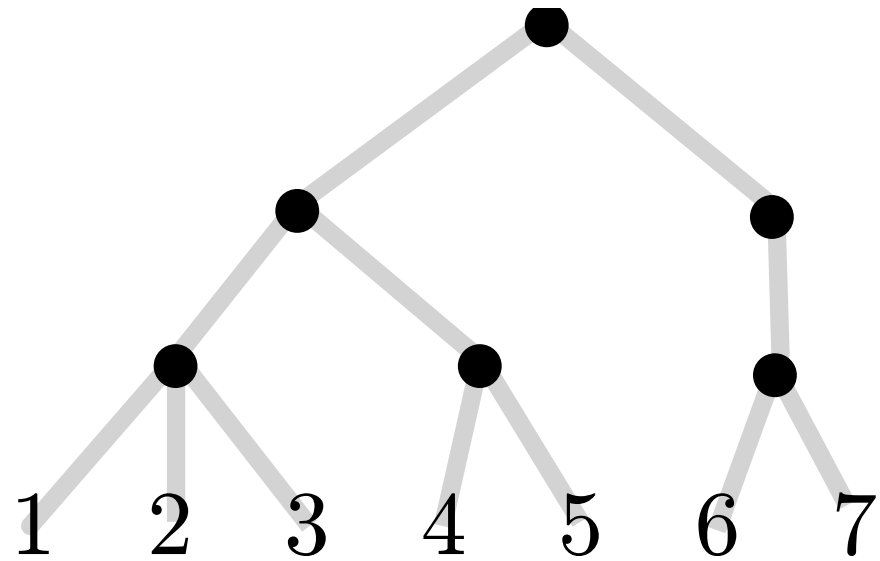
Motivation

- True hierarchies
- Parameter tying
- Visualisation and interpretability



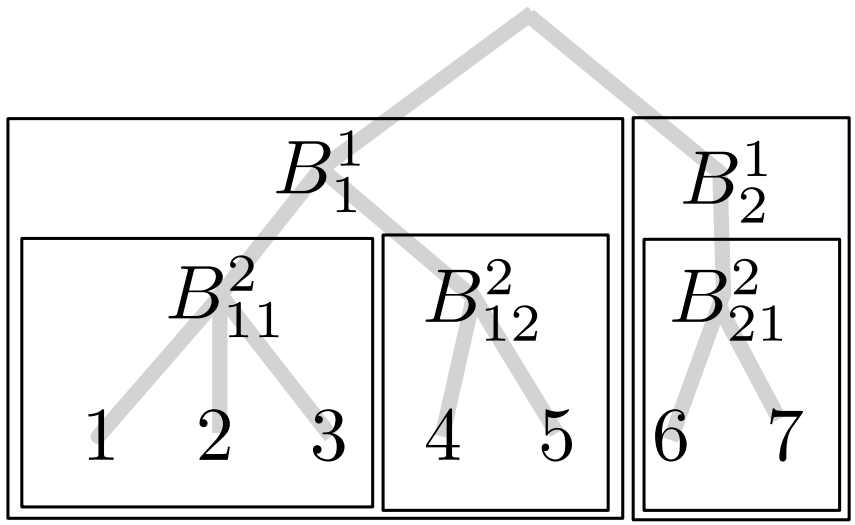
Trees

- For statistical applications we are usually interested in trees with *labelled* leaves.
- Mathematically we can think of this as a *hierarchical partition* or a *acyclic graph*.
- Usually we distinguish the root node.
- e.g. for $\{\{\{1,2,3\},\{4,5\}\},\{\{6,7\}\}\}$ the tree is



Nested CRP

- Distribution over hierarchical partitions
- Denote the K blocks in the first level as $\{B_k^1 : k = 1, ..., K\}$
- Partition these blocks with independent CRPs
- Denote the partition of B_k^1 as $\{B_{kl}^2 : l = 1, ..., K_k\}$
- Recurse for S iterations, forming a S deep hierarchy



- Used to define a *hierarchical topic model*, see Blei et al. (2010)

nCRP: when to stop?

Two approaches

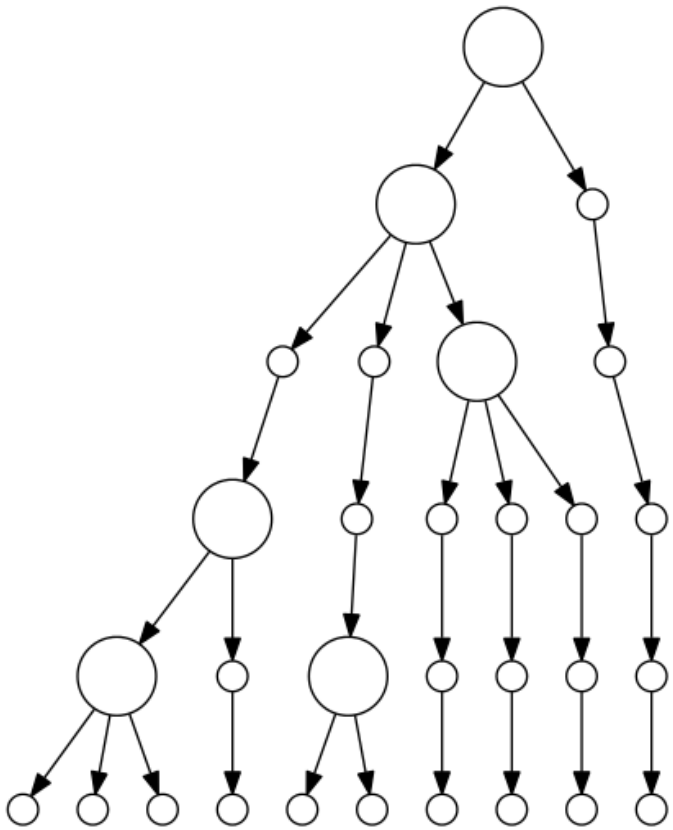
- Use a finite depth S
- Work with the infinitely deep tree

In the later case we can either

- Augment with a per node probability of stopping, e.g. Adams et al. (2009)
- Integrate over chains of infinite length, e.g. Steinhardt and Ghahramani (2012)

The Dirichlet diffusion tree

- The Dirichlet diffusion tree (DDT, Neal (2003)) is the *continuum limit* of the nCRP
- Embed an S-deep nCRP in continuous time $\mathbb{R}^+$ such that the "time" of level i is $i/\sqrt{S}$
- Set the concentration parameter of each CRP to $\theta = 1/\sqrt{S}$
- Take the limit as $S \to \infty$
- Ignore degree two nodes (apart from the root)
- What does this process look like?



The Dirichlet diffusion tree

- Item 1 sits at its own table for all time $\mathbb{R}^+$

For item i

- sits at the same table as all the other previous items at time $t = 0$
- In time interval $[t, t + dt]$, splits off to start its own table with probability dt/m, where m is the number of previous items at this table (at time t)
- If a previous item started a new table (this is a "branch point") then choose the two tables with probability proportion to the number of items that previously went each way

The Dirichlet diffusion tree

- While the nCRP allowed arbitrary branching, the DDT only has binary branching events: this is addressed by the Pitman Yor diffusion tree (K and Ghahramani, 2011)
- Branches now have associated *branch lengths*: these came from chains in the nCRP
- Items will "diverge" to form a singleton table at some finite t almost surely
- Draw this

As a model for data

- We have a prior over trees with unbounded depth and width. However, difficult to use with unbounded divergence times.
- Solution: transform the branch times t according to $t' = A^{-1}(t)$ where $A^{-1} : \mathbb{R}^+ \to [0, 1]$ and is strictly increasing, e.g. $A^{-1}(t) = 1 - e^{-t/c}, A(t') = -c \log(1 - t')$
- Equivalent to changing the probability of divergence in $[t, t + dt]$ to $a(t)dt/m$ where $a(t) = A'(t) = c/(1 - t)$.

Diffusion on the tree

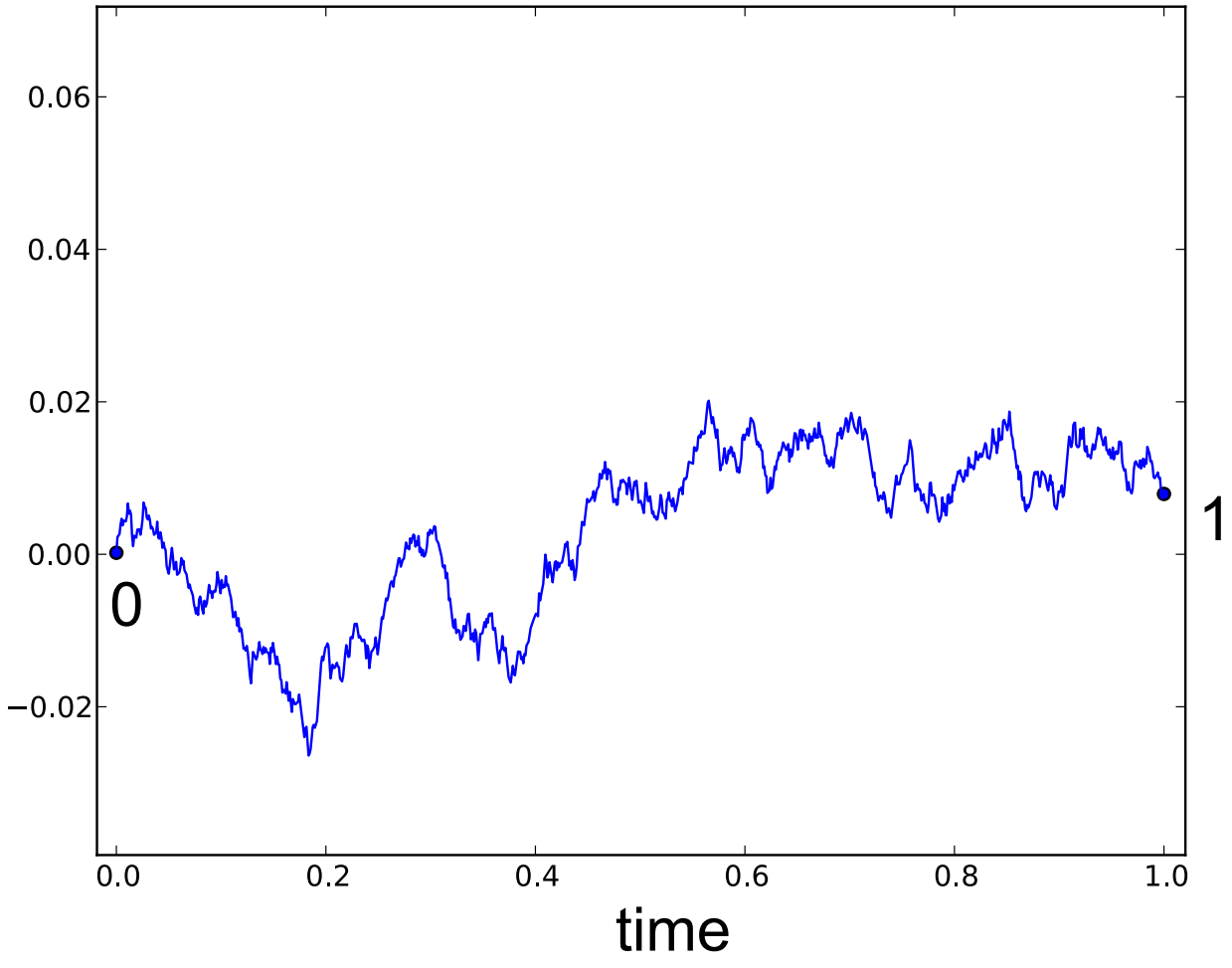
- We will model data x at the leaves using a diffusion process on the tree
- Interpretation of the tree structure as a graphical model

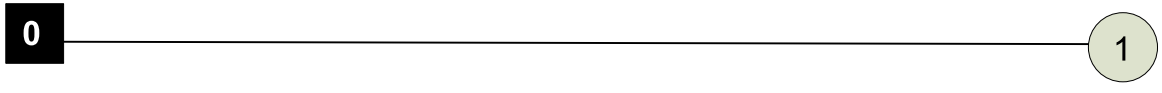
$$P(x_{\text{child}}|x_{\text{parent}}) = k(x_{\text{child}}|x_{\text{parent}}, t_{\text{child}} - t_{\text{parent}})$$

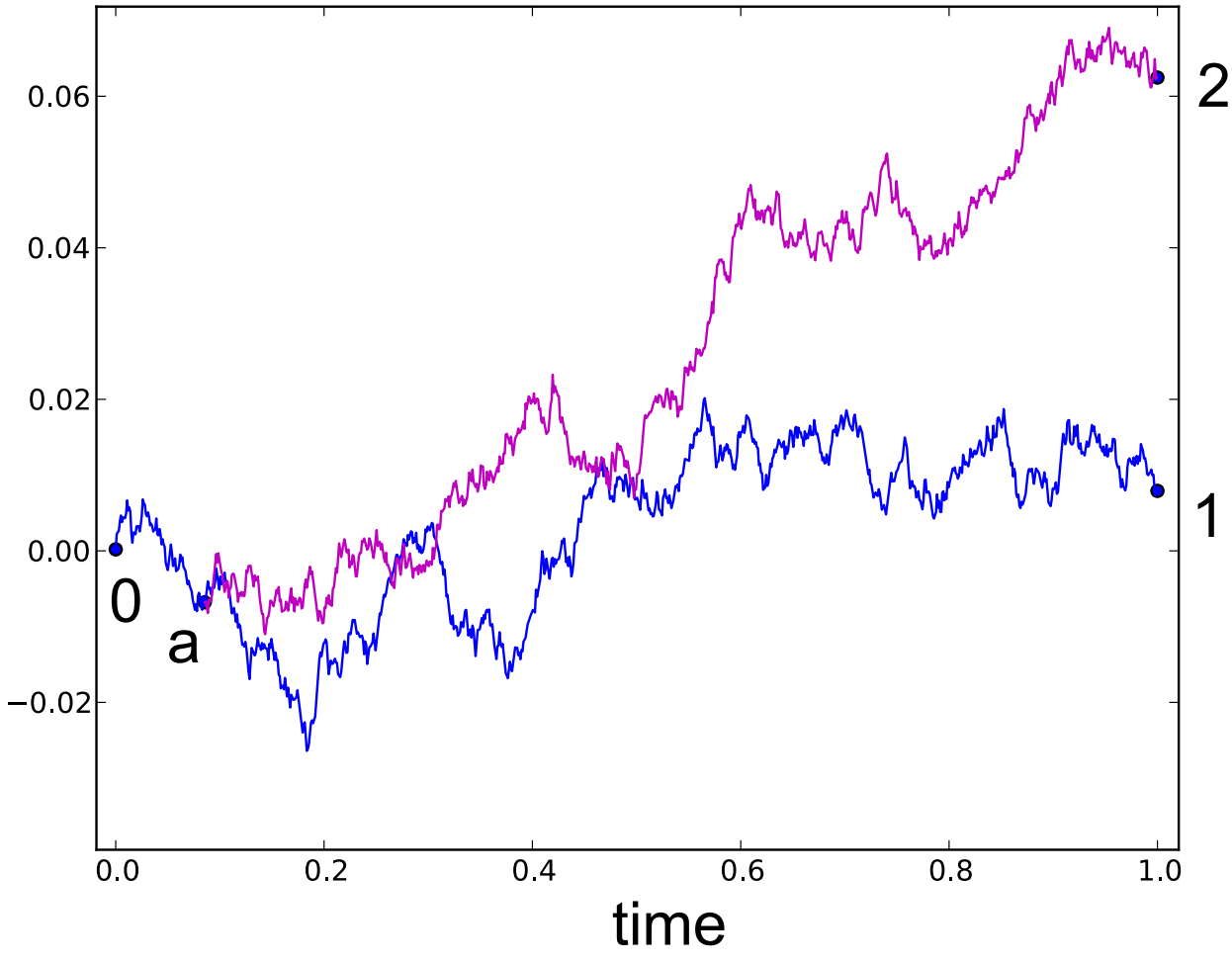
- We will focus on Brownian (Gaussian) diffusion:

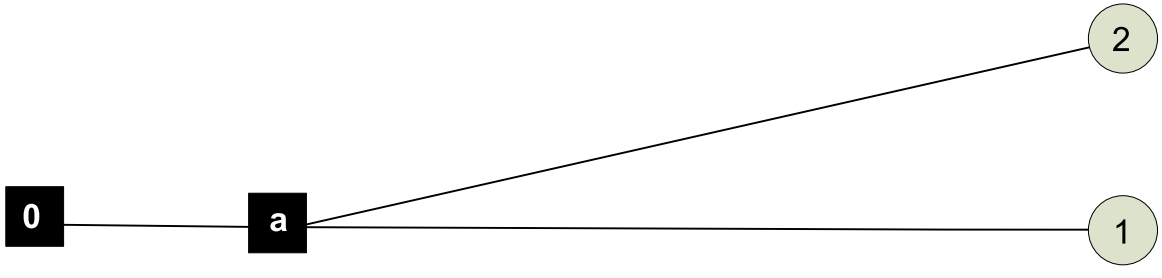
$$k(x_{\text{child}}|x_{\text{parent}}, t_{\text{child}} - t_{\text{parent}}) = N(x_{\text{child}}; x_{\text{parent}}, t_{\text{child}} - t_{\text{parent}})$$

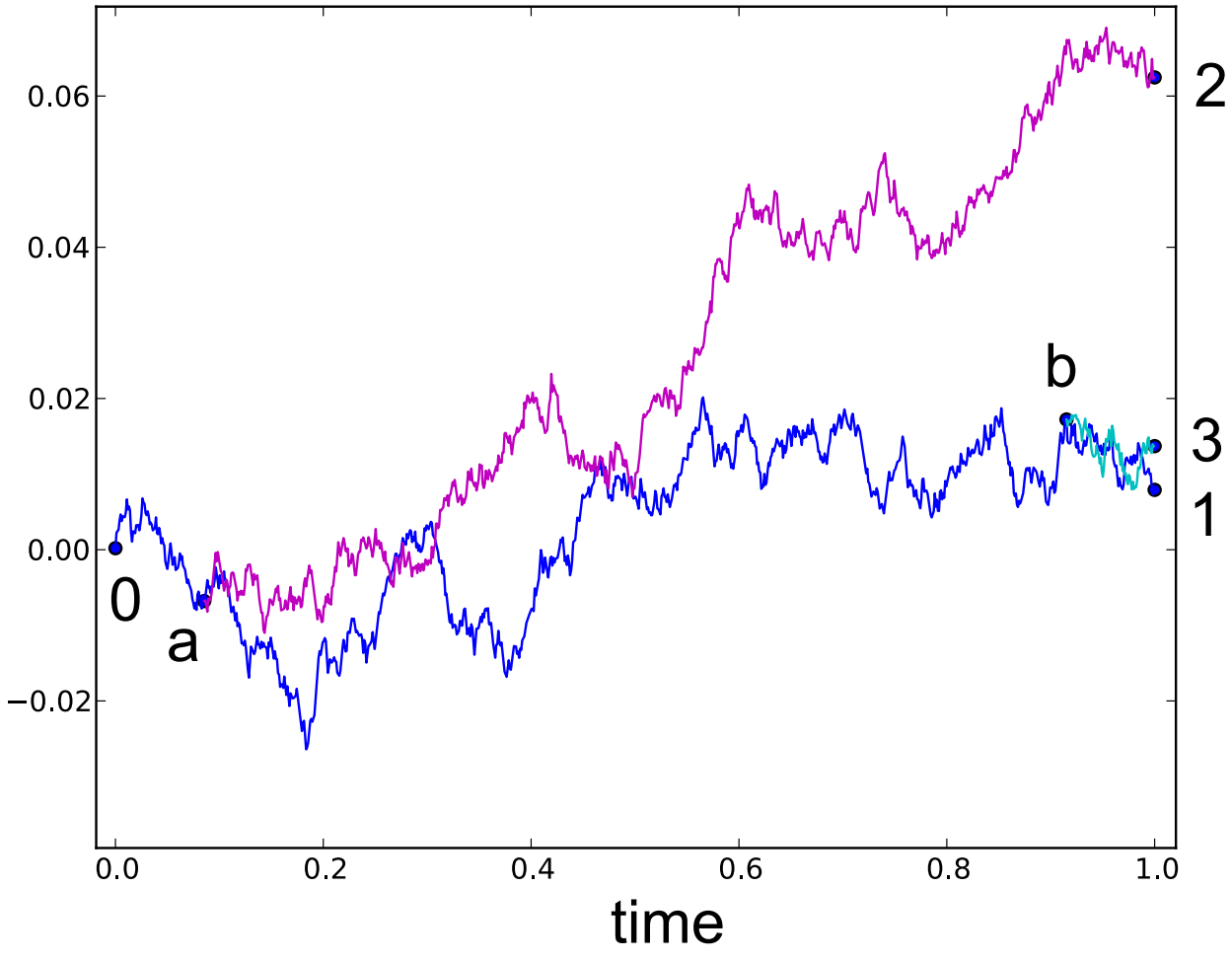
- This allows the marginal likelihood $P(x_{\text{leaves}}|\text{tree})$ to be calculated using a single upwards sweep of message passing (belief propagation)

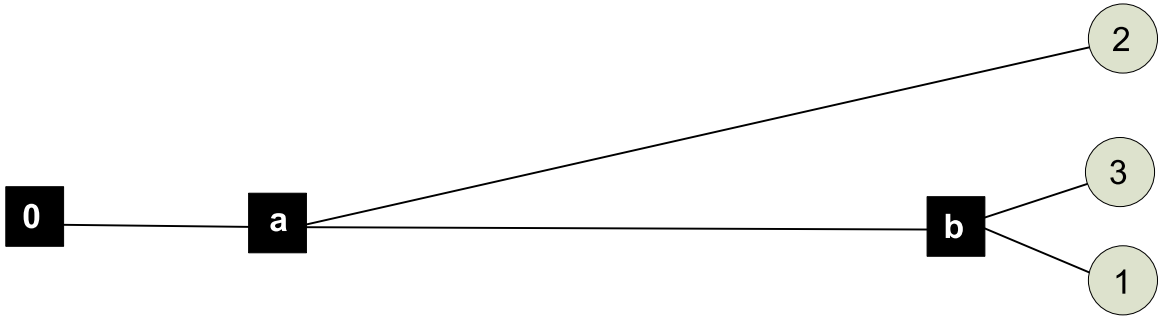

0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
1


0
1


0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
a
2
1


0
a
2
1


0.06
0.04
0.02
0.00
−0.02
0.0
0.2
0.4
0.6
0.8
1.0
time
0
a
b
2
3
1


0
a
b
2
3
1

Properties

- Exchangeable (we can show this by explicit calculational of the probability of the tree structure, divergence times and locations), see proof in Neal (2003)
- Projective by its sequential construction
- Positive support for any binary tree

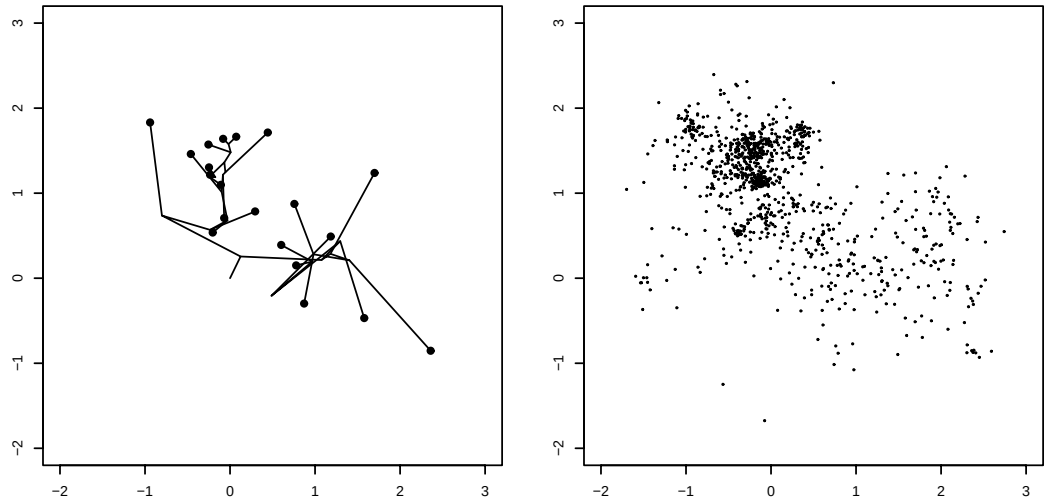


Figure 3. *Generation of a two-dimensional data set from the Dirichlet diffusion tree prior with* $\sigma = 1$ *and* $a(t) = 1/(1-t)$*. The plot on the left shows the first twenty data points generated, along with the underlying tree structure. The right plot shows 1000 data points obtained by continuing the procedure beyond these twenty points.*

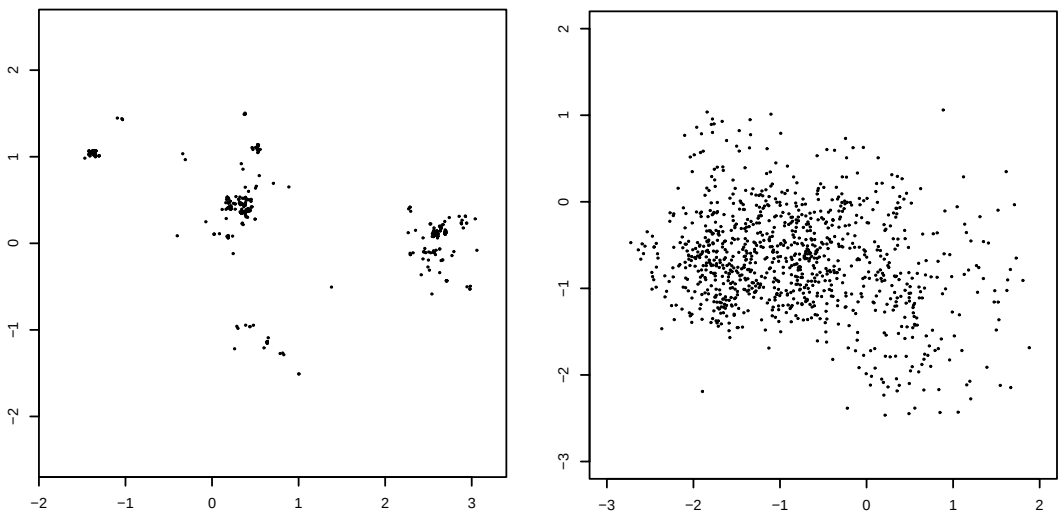


Figure 4. *Two data sets of 1000 points drawn from Dirichlet diffusion tree priors with* $\sigma = 1$*. For the data set on the left, the divergence function used was* $a(t) = (1/4)/(1-t)$*. For the data set on the right,* $a(t) = (3/2)/(1-t)$*.*

Kingman's coalescent

- The DDT is an example of a Gibbs *fragmentation* tree (McCullagh et al., 2008)
- We can also construct trees using *coagulation* processes
- Kingman's coalescent (KC Kingman, 1982): iteratively *merge* subtrees, starting with all leaves in their own subtrees.
- Can also be derived as the continuum limit of a population genetics model (the Wright-Fisher model) of large populations of haploid individuals (i.e. only one parent)

Kingman's coalescent

- Subtrees merge independently with rate 1: for m subtrees the time until the next merge is $Exp(m(m-1)/2)$
- Used in a similar manner to the DDT for hierarchical clustering in Teh et al. (2008)

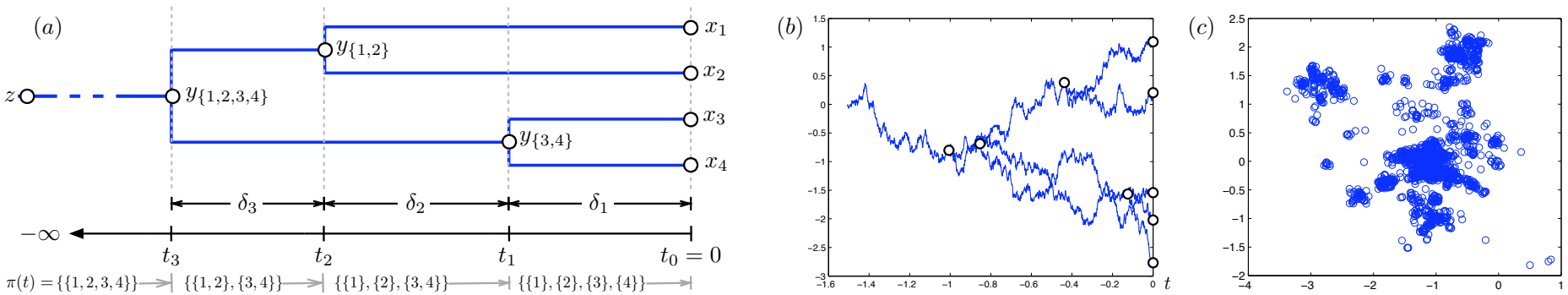
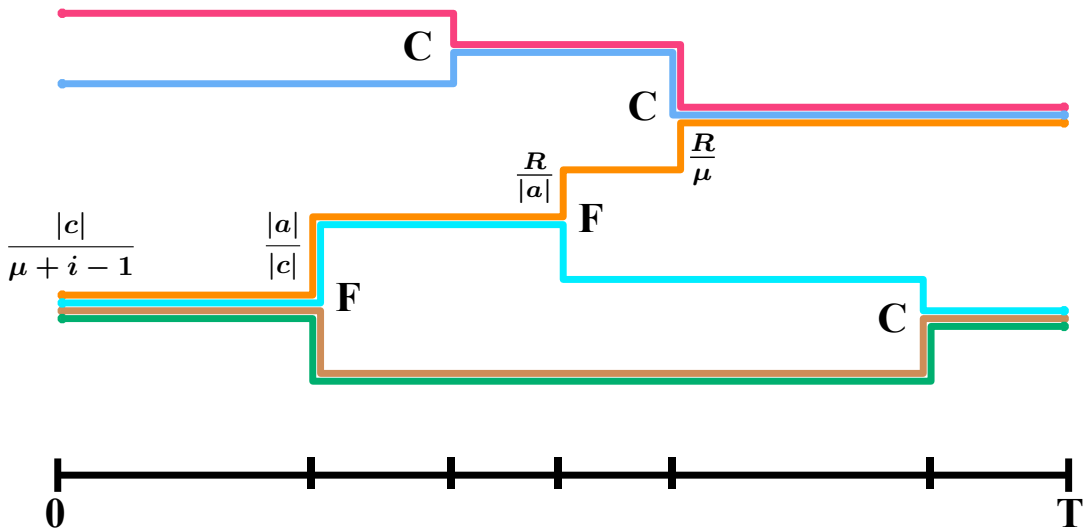


Figure 1: (a) Variables describing the n-coalescent. (b) Sample path from a Brownian diffusion coalescent process in 1D, circles are coalescent points. (c) Sample observed points from same in 2D, notice the hierarchically clustered nature of the points.

Fragmentation-coagulation processes

- The DDT and KC are dual in the following sense: start with a CRP distributed partition, run DDT for dt, then KC for dt, and the result is still CRP distributed with the same parameters
- Used to construct the fragmentation-coagulation (FC) process: a reversible, Markov partition-valued process with the CRP as its stationary distribution
- Applied to modelling genetic variation (Teh et al., 2011)

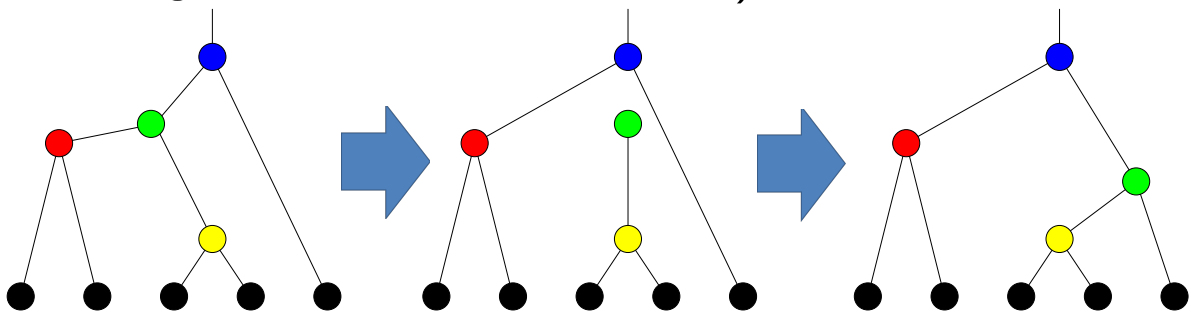


Tree models I didn't talk about

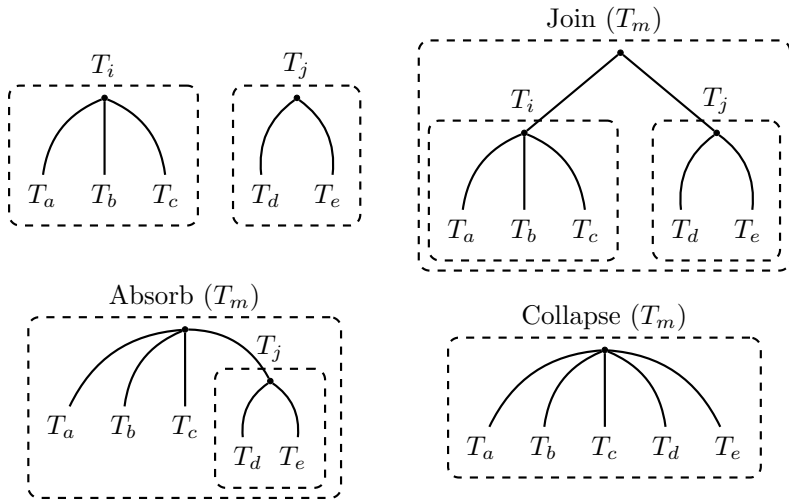
- λ-coalescents: broad class generalising Kingman's coalescent (Pitman, 1999)
- Polya trees: split the unit interval recursively (binary splits) (Lavine, 1992; Mauldin et al., 1992), used for univariate density modelling
- Tree-structured stick breaking: a stick breaking representation of the nested CRP. Uses per node stopping probabilities (Adams et al., 2009)
- Time marginalised KC (Boyles and Welling, 2012)
- No prior (implicitly uniform)! Some work does ML estimation of trees while integrating over other variables (Heller and Ghahramani, 2005; Blundell et al., 2010)

Inference

- Detaching and reattaching subtrees (Neal, 2003; Boyles and Welling, 2012; K et al., 2011)



- Sequential Monte Carlo (Bouchard-Côté et al., 2012; Gorur and Teh, 2008)
- Greedy agglomerative methods (Heller and Ghahramani, 2005; Blundell et al., 2010)



Conclusions

- Exchangeability, projectivity and support are key characteristics to consider when designing models
- Proving consistency and convergence rates is an active area of research
- There are a lot of priors over the space of trees with interesting relationsips between them
- Rich probability literature on many of these processes

Bibliography I

Adams, R., Ghahramani, Z., and Jordan, M. (2009). Tree-Structured Stick Breaking Processes for Hierarchical Modeling. *Statistics*, (1):1–3.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57.

Blundell, C., Teh, Y. W., and Heller, K. A. (2010). Bayesian rose trees. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*.

Bouchard-Côté, A., Sankararaman, S., and Jordan, M. I. (2012). Phylogenetic inference via sequential monte carlo. *Systematic biology*, 61(4):579–593.

Boyles, L. and Welling, M. (2012). The time-marginalized coalescent prior for hierarchical clustering. In *Advances in Neural Information Processing Systems 25*, pages 2978–2986.

Gorur, D. and Teh, Y. (2008). A Sequential Monte Carlo Algorithm for Coalescent Clustering. *gatsby.ucl.ac.uk*.

Bibliography II

Heller, K. A. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the 22nd International Conference on Machine learning*, page 304.

K, D. A., Gael, J. V., and Ghahramani, Z. (2011). Message passing algorithms for Dirichlet diffusion trees. In *Proceedings of the 28th Annual International Conference on Machine Learning*.

K, D. A. and Ghahramani, Z. (2011). Pitman-Yor diffusion trees. In *The 28th Conference on Uncertainty in Artificial Intelligence (to appear)*.

Kingman, J. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235 – 248.

Lavine, M. (1992). Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics*, pages 1222–1235.

Mauldin, R. D., Sudderth, W. D., and Williams, S. (1992). Polya trees and random distributions. *The Annals of Statistics*, pages 1203–1221.

McCullagh, P., Pitman, J., and Winkel, M. (2008). Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002.

Neal, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629.

Bibliography III

Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902.

Steinhardt, J. and Ghahramani, Z. (2012). Flexible martingale priors for deep hierarchies. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Teh, Y. and Orbanz, P. (2011). Nips tutorial: Bayesian nonparametrics. In *NIPS*.

Teh, Y. W., Blundell, C., and Elliott, L. T. (2011). Modelling genetic variations with fragmentation-coagulation processes. In *Advances in Neural Information Processing Systems (NIPS)*.

Teh, Y. W., Daumé III, H., and Roy, D. M. (2008). Bayesian agglomerative clustering with coalescents. *Advances in Neural Information Processing Systems*, 20.