Computer Science @ RIT
B. Thomas Golisano College of Computing & Information Sciences

Pitman Yor Diffusion Trees for Bayesian Hierarchical Clustering

David A. Knowles and Zoubin Ghahramani

IEEE Transactions on Pattern Analysis and Machine Intelligence,
Vol. 37, No. 2, February 2015

Presented by Robert McCartney and Jie Yuan

Pitman-Yor Diffusion Tree

Bayesian nonparametric prior over tree structures with arbitrary branching structure at each branch point to create an exchangeable distribution over data points

What?

- Nonparametric methods make no (or fewer) assumptions on the underlying probability distributions
- For instance, how many clusters are there?
- With k-means this is a parameter to the model
- Nonparametric methods want to let the data decide this

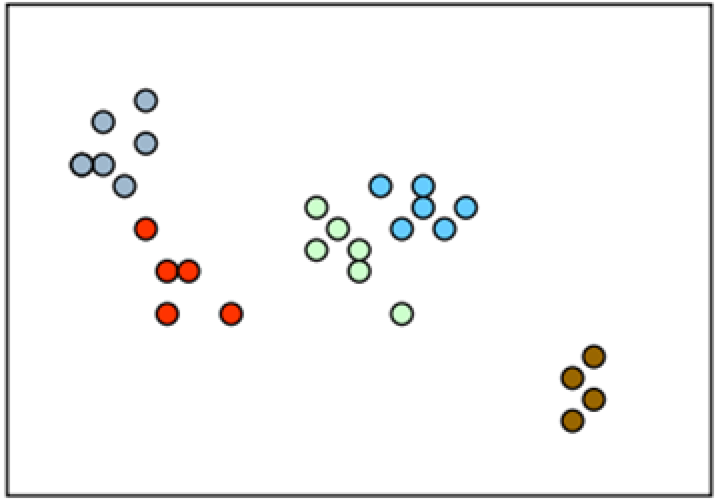


Figure from https://www.cs.cmu.edu/~kbe/dp_tutorial.pdf

Ok, go on..

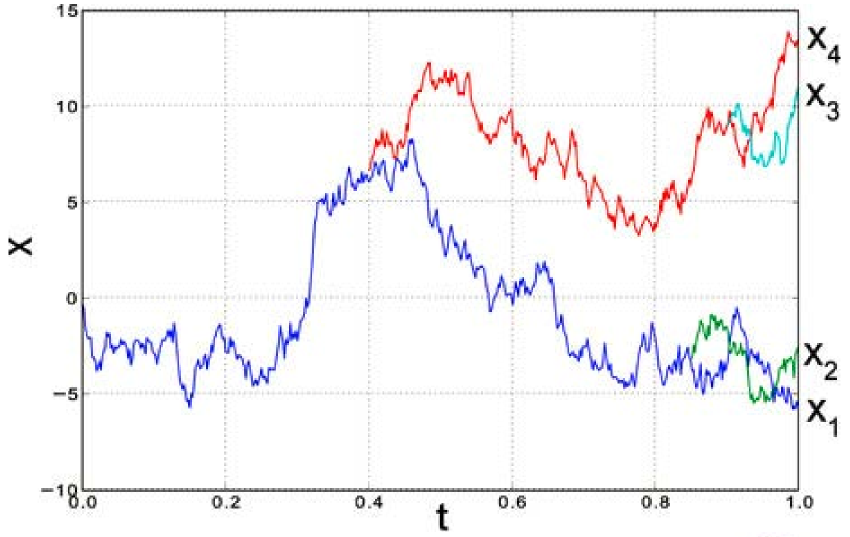
- Tree structures are used to hierarchically cluster the data
- They give interpretable results in data exploration stages
- They provide density estimations
- This is paper extends previous methods related to Dirichlet Diffusion Trees

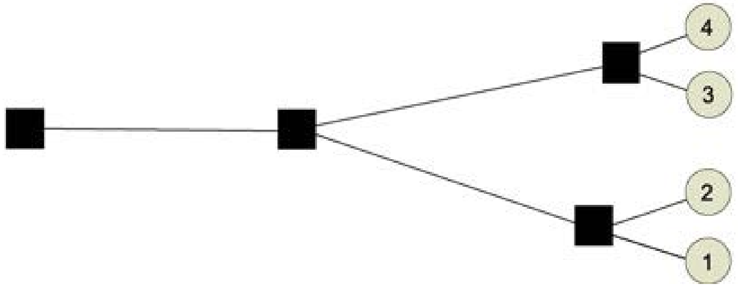
Dirichlet Diffusion Trees

- DDTs are like playing plinko
- Points start out following the path traced by previous points
- At each interval $\mathrm{dt} \in [0,1]$ the probability of diverging from current path with *m* datapoints having traveled it is:

$$\frac{a(t)dt}{m}, where\ a(t) = \frac{c}{1-t}$$

- Brownian motion with Gaussian distributions model the final destination of points

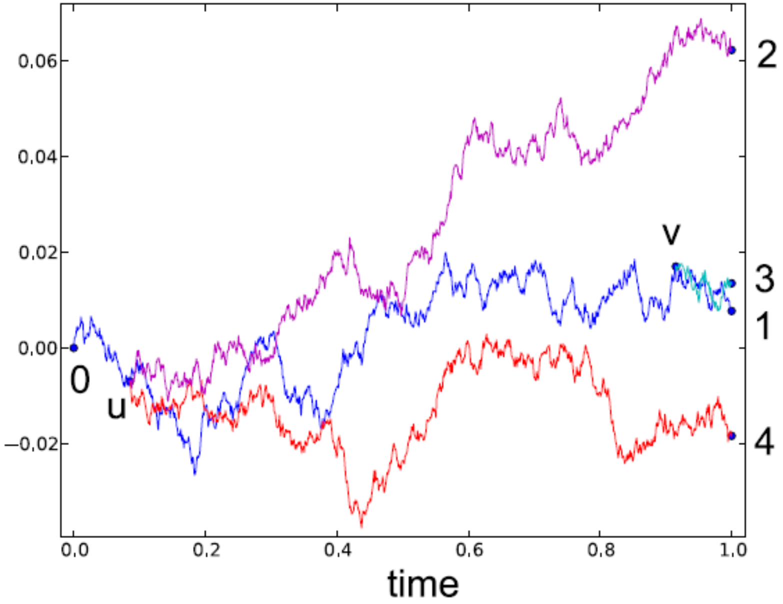


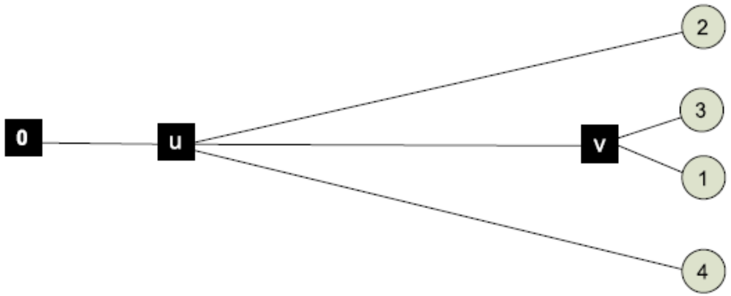


This paper - PYDT

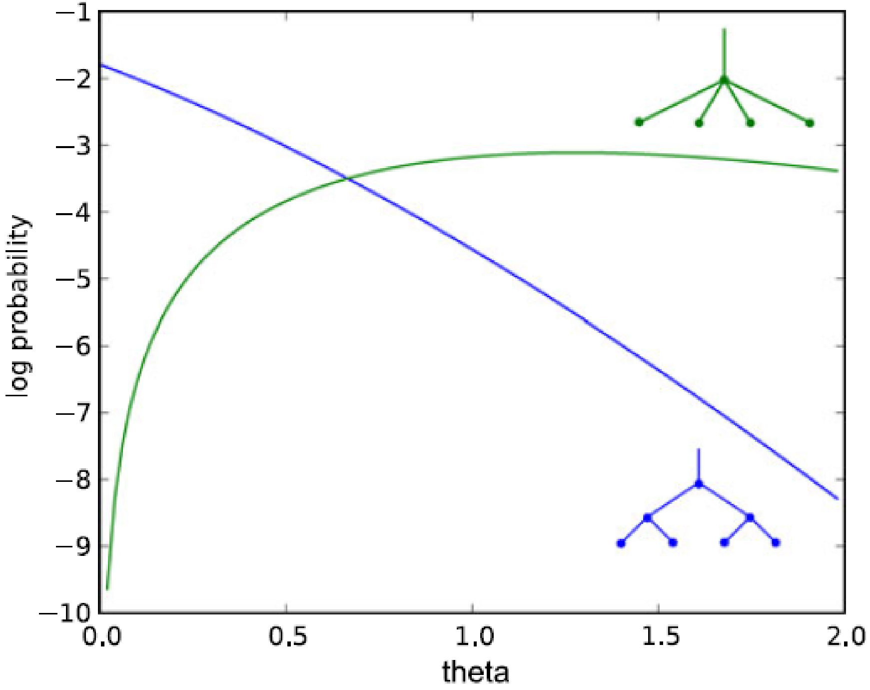
- <u>Arbitrary branching</u> means instead of binary there can be K branches
- Probability of following an existing path with n_k samples is $\frac{n_k - \alpha}{m+\theta}$ and diverging is $\frac{\theta + \alpha K}{m+\theta}$
- This sums to 1 since $\sum n_k = m$

- <u>Exchangeable distribution</u> means the probability of obtaining any given tree structure is invariant to reordering the data points
- This was proven in the paper





Recovering DDTs



- When $\theta = \alpha = 0$ this becomes a DDT
- θ determines branching behavior

You said no parameters

- The hyperparameters themselves $(\alpha, c, \theta, \sigma)$ are given prior distributions
- Markov Chain Monte Carlo is used to detach and reattach subtrees for thousands of iterations to maximize the tree’s marginal probability based off of the prior probabilities of the data
- Final tree is a density model of the joint distribution over the data and hyperparameters

Comparison to DDT

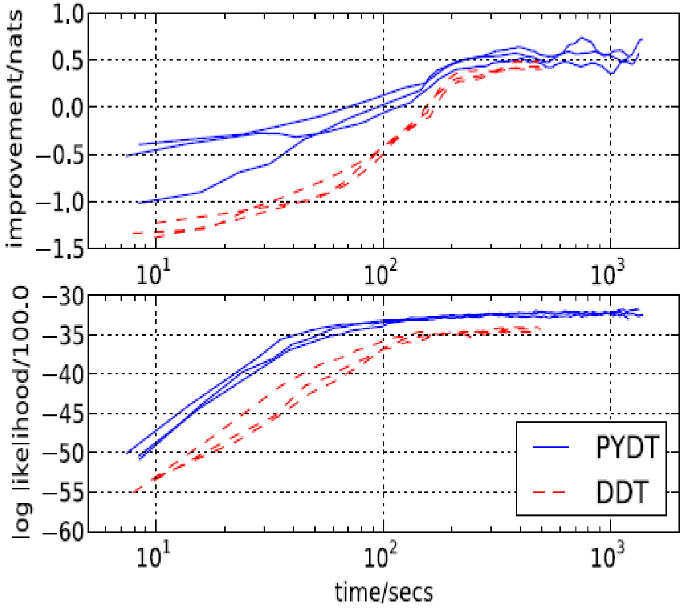


Fig. 11. Density modelling of the $D = 10, N = 200$ macaque skull measurement data set of [1]. *Top*: Improvement in test predictive likelihood compared to a kernel density estimate. *Bottom*: Marginal likelihood of current tree. The shared x-axis is computation time in seconds.

1. Moderate improvement in predictive performance

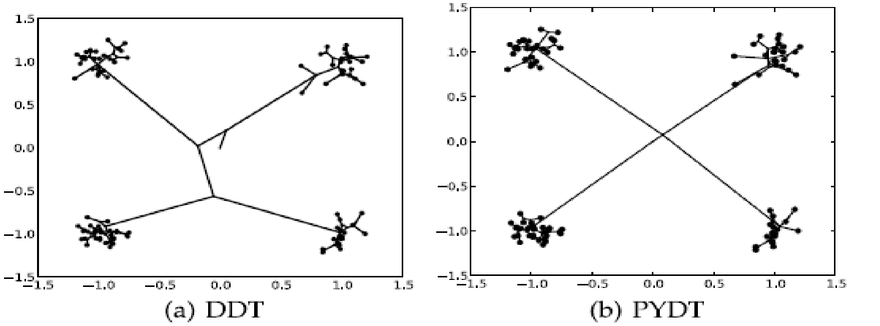


Fig. 10. Optimal trees learnt by the greedy EM algorithm for the DDT and PYDT on a synethic data set with $D = 2, N = 100$.

2. Improved interpretability through logical hierarchies

Questions

Backup slides

Dirichlet Process

Draw distribution G from DP(G0, alpha)
Draw observations x independently from G

xi|thetai ~ F(thetai)
thetai|G ~ G
G|alpha, H ~ DP(alpha, H)

(G(A1), … G(Ak) ~ Dir(alpha*H(A1), …
alpha*H(Ak))

Pitman Yor Diffusion Tree

Generalizes Dirichlet Diffusion Tree to allow arbitrary branching

probability of breaking from previous path is now a(t)Gamma(m-alpha)dt/Gamma(m+1+theta)

theta is concentration parameter (Dirichlet alpha)

When alpha = 0, Pitman Yor process reduces to CRP, but the more occupied tables, the more likely people will join new tables; tables with fewer people have less chance of getting more

At every branch point, go down existing path with prob (n_k - alpha)/(m + theta), but also diverge at branch point with prob (theta + alpha*K)/(m + theta) where K is existing #

Dirichlet Process (Chinese restaurant process)
- nonparametric selection of clusters given data

Dirichlet Diffusion Tree - generalization of
Dirichlet process (similar to CRP)

Pitman Yor Diffusion Tree (this paper) -
generalization of Dirichlet Diffusion Tree
(two parameter Chinese restaurant process vs

Chinese Restaurant Process

Each customer sits at a new table (K+1) with probability alpha/(alpha+n-1)

Customer sits at table k with probability n_k/(alpha+n-1)

each table represents a base distribution with some parameters

number of tables grows logarithmically with the data

http://blog.echen.me/2012/03/20/infinite-mixture-models-with-nonparametric-bayes-and-the-dirichlet-process/
http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/dp.pdf