A Brief Overview of Nonparametric Bayesian Models

NIPS 2009 Workshop

Zoubin Ghahramani[1]

Department of Engineering
University of Cambridge, UK

zoubin@eng.cam.ac.uk
http://learning.eng.cam.ac.uk/zoubin

[1] Also at Machine Learning Department, CMU

Parametric vs Nonparametric Models

- *Parametric models* assume some finite set of parameters θ. Given the parameters, future predictions, x, are independent of the observed data, $\mathcal{D}$:

$$P(x|\theta, \mathcal{D}) = P(x|\theta)$$

  therefore θ capture everything there is to know about the data.

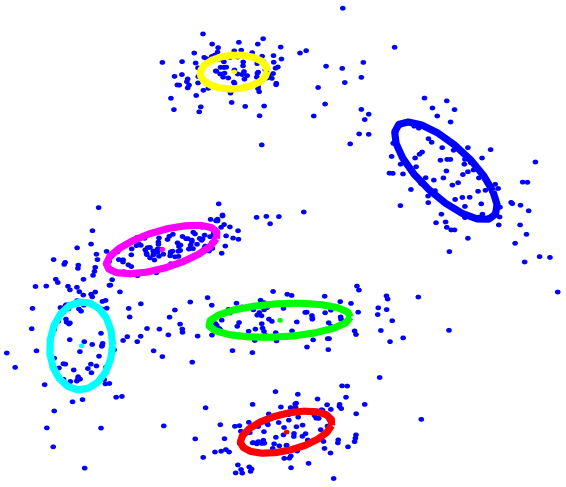
- So the complexity of the model is bounded even if the amount of data is unbounded. This makes them not very flexible.

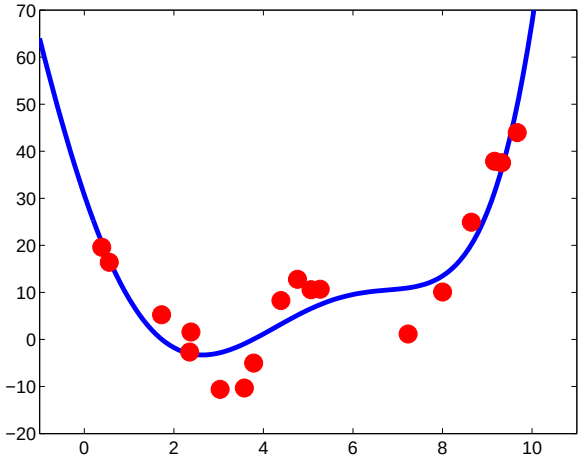
- *Non-parametric models* assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an *infinite dimensional* θ. Usually we think of θ as a *function*.

- The amount of information that θ can capture about the data $\mathcal{D}$ can grow as the amount of data grows. This makes them more flexible.

Why?

- flexibility
- better predictive performance
- more realistic





Outline

Bayesian nonparametrics has many uses.

Some modelling goals and *examples* of associated nonparametric Bayesian models:

Modelling Goal	Example process
Distributions on functions	Gaussian process
Distributions on distributions	Dirichlet process
	Polya Tree
Clustering	Chinese restaurant process
	Pitman-Yor process
Hierarchical clustering	Dirichlet diffusion tree
	Kingman's coalescent
Sparse latent feature models	Indian buffet processes
Survival analysis	Beta processes
Distributions on measures	Completely random measures
...	...

Gaussian Processes

A Gaussian process defines a distribution $P(f)$ on functions, f, where f is a function mapping some input space $\mathcal{X}$ to $\Re$.

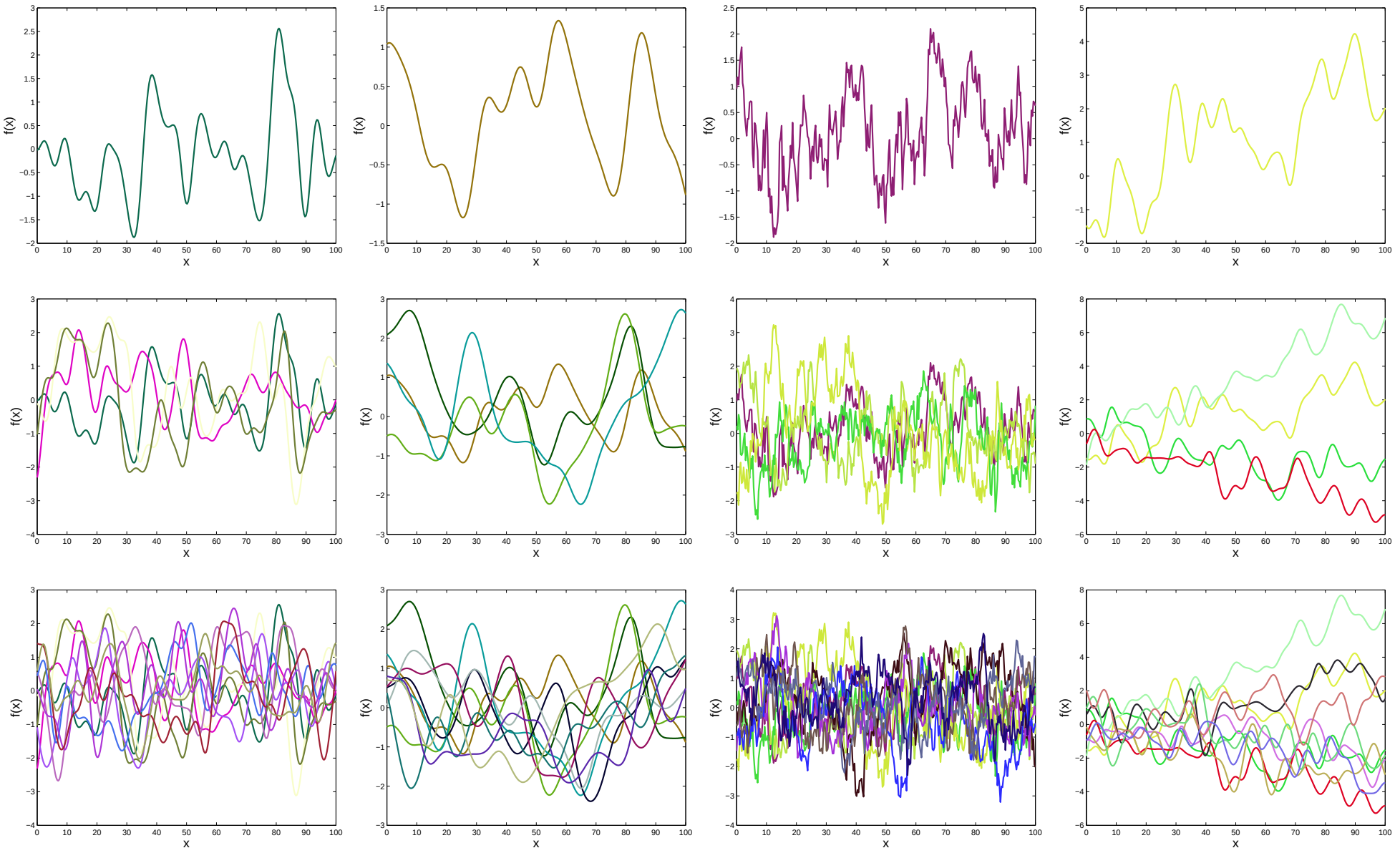
$$f : \mathcal{X} \to \Re.$$

Let $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))$ be an n-dimensional vector of function values evaluated at n points $x_i \in \mathcal{X}$. Note $\mathbf{f}$ is a random variable.

Definition: $P(f)$ is a Gaussian process if for *any* finite subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the marginal distribution on that finite subset $P(\mathbf{f})$ has a multivariate Gaussian distribution, and all these marginals are coherent.

We can think of GPs as the *extension* of multivariate Gaussians to distributions on functions.

Samples from Gaussian processes with different $c(x, x')$



Dirichlet Processes

- Gaussian processes define a distribution on functions

$$f \sim \mathsf{GP}(\cdot|\mu, c)$$

  where μ is the mean function and c is the covariance function.
  We can think of GPs as "infinite-dimensional" Gaussians

- Dirichlet processes define a distribution on distributions (a measure on measures)

$$G \sim \mathsf{DP}(\cdot|G_0, \alpha)$$

  where $\alpha > 0$ is a scaling parameter, and G_0 is the base measure.
  We can think of DPs as "infinite-dimensional" Dirichlet distributions.

Note that both f and G are infinite dimensional objects.

Dirichlet Process

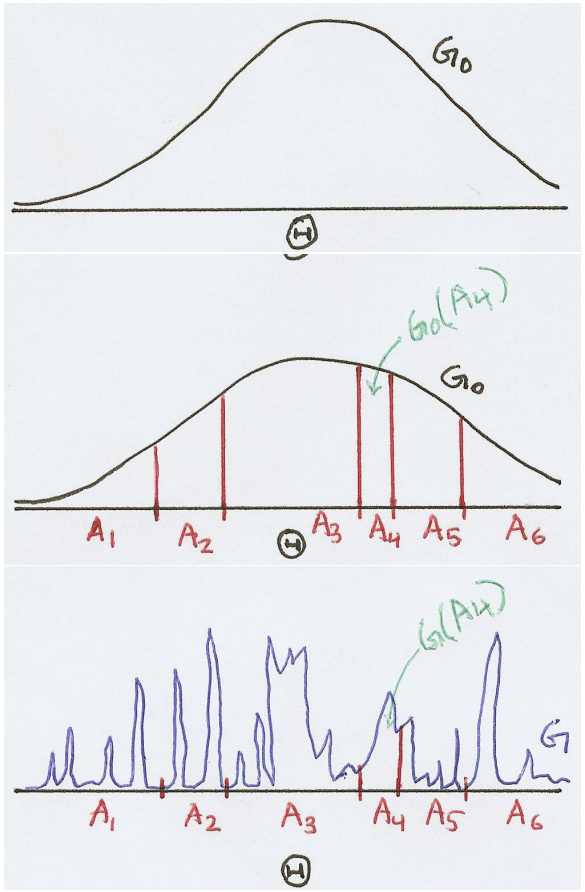
Let Θ be a measurable space, G_0 be a probability measure on Θ, and α a positive real number.

For all $(A_1, \dots A_K)$ finite partitions of Θ,

$$G \sim \mathsf{DP}(\cdot | G_0, \alpha)$$

means that

$$(G(A_1), \dots, G(A_K)) \sim \mathsf{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_K))$$



(Ferguson, 1973)

Dirichlet Distribution

The Dirichlet distribution is a distribution on the K-dim probability simplex.

Let $\mathbf{p}$ be a K-dimensional vector s.t. $\forall j : p_j \geq 0$ and $\sum_{j=1}^{K} p_j = 1$

$$P(\mathbf{p}|\boldsymbol{\alpha}) = \text{Dir}(\alpha_1, \ldots, \alpha_K) \stackrel{\text{def}}{=} \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}$$

where the first term is a normalization constant[2] and $E(p_j) = \alpha_j / (\sum_k \alpha_k)$

The Dirichlet is conjugate to the multinomial distribution. Let

$$c|\mathbf{p} \sim \text{Multinomial}(\cdot|\mathbf{p})$$

That is, $P(c = j|\mathbf{p}) = p_j$. Then the posterior is also Dirichlet:

$$P(\mathbf{p}|c = j, \boldsymbol{\alpha}) = \frac{P(c = j|\mathbf{p})P(\mathbf{p}|\boldsymbol{\alpha})}{P(c = j|\boldsymbol{\alpha})} = \text{Dir}(\boldsymbol{\alpha}')$$

where $\alpha'_j = \alpha_j + 1$, and $\forall \ell \neq j : \alpha'_\ell = \alpha_\ell$

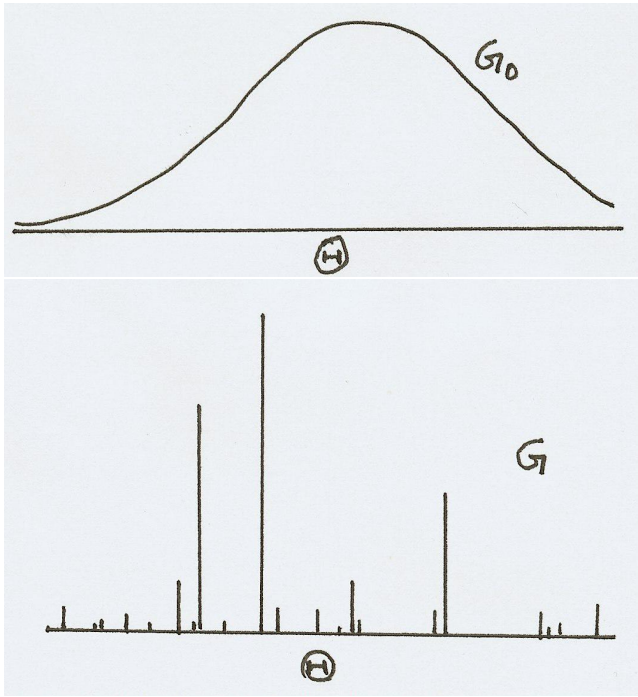
[2] $\Gamma(x) = (x-1)\Gamma(x-1) = \int_0^\infty t^{x-1} e^{-t} dt$. For integer n, $\Gamma(n) = (n-1)!$

Dirichlet Process

$G \sim \text{DP}(\cdot|G_0, \alpha)$ OK, but what does it look like?

Samples from a DP are discrete with probability one:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$



where $\delta_{\theta_k}(\cdot)$ is a Dirac delta at θ_k, and $\theta_k \sim G_0(\cdot)$.

Note: $E(G) = G_0$

As $\alpha \to \infty$, G looks more "like" G_0.

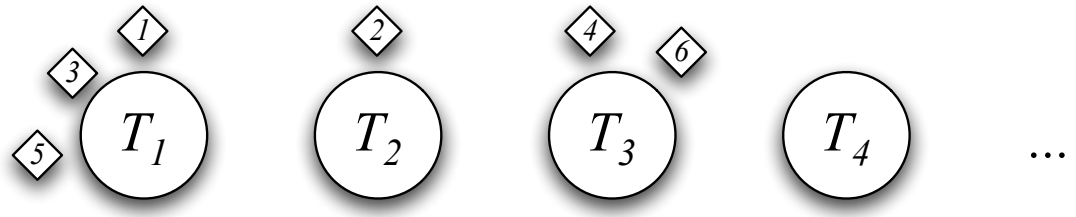
Relationship between DPs and CRPs

- DP is a distribution on distributions
- DP results in discrete distributions, so if you draw n points you are likely to get repeated values
- A DP induces a partitioning of the n points
  e.g. $(1\,3\,4)\,(2\,5) \Leftrightarrow \theta_1 = \theta_3 = \theta_4 \neq \theta_2 = \theta_5$
- Chinese Restaurant Process (CRP) defines the corresponding distribution on partitions
- Although the CRP is a sequential process, the distribution on $\theta_1, \ldots, \theta_n$ is exchangeable (i.e. invariant to permuting the indices of the θs): e.g.

$$P(\theta_1, \theta_2, \theta_3, \theta_4) = P(\theta_2, \theta_4, \theta_3, \theta_1)$$

Chinese Restaurant Process

The CRP generates samples from the distribution on partitions induced by a DPM.



Generating from a CRP:

customer 1 enters the restaurant and sits at table 1.
$K = 1$, $n = 1$, $n_1 = 1$
for $n = 2, \ldots,$

customer n sits at table $\begin{cases} k & \text{with prob } \frac{n_k}{n-1+\alpha} & \text{for } k = 1 \ldots K \\ K+1 & \text{with prob } \frac{\alpha}{n-1+\alpha} & \text{(new table)} \end{cases}$

if new table was chosen **then** $K \leftarrow K + 1$ **endif**
endfor

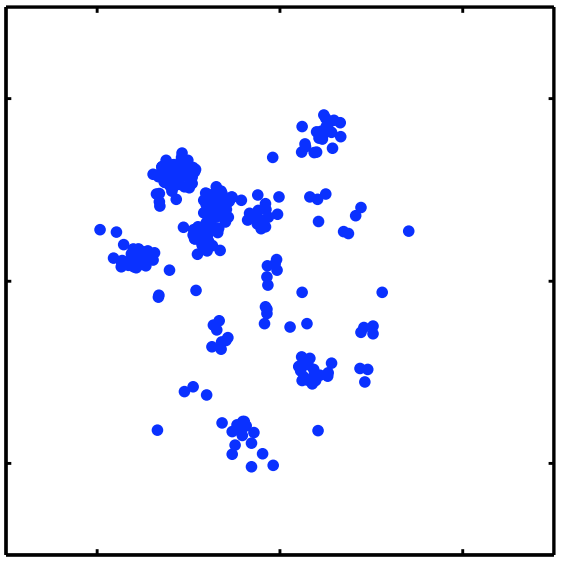
"Rich get richer" property.

(Aldous 1985; Pitman 2002)

Dirichlet Processes: Big Picture

There are many ways to derive the Dirichlet Process:

- Dirichlet distribution
- Urn model
- Chinese restaurant process
- Stick breaking
- Gamma process



DP: distribution on distributions

Dirichlet process mixture (DPM): a mixture model with infinitely many components where parameters of each component are drawn from a DP. Useful for clustering; assignments of points to clusters follows a CRP.

Hierarchical Clustering

Dirichlet Diffusion Trees (DFT)

(Neal, 2001)

In a DPM, parameters of one mixture component are independent of another components – this lack of structure is potentially undesirable.

A DFT is a generalization of DPMs with hierarchical structure between components.

To generate from a DFT, we will consider θ taking a random walk according to a Brownian motion Gaussian diffusion process.

- $\theta_1(t) \sim$ Gaussian diffusion process starting at origin ($\theta_1(0) = 0$) for unit time.
- $\theta_2(t)$, also starts at the origin and follows θ_1 but diverges at some time τ_d, at which point the path followed by θ_2 becomes independent of θ_1's path.
- $a(t)$ is a divergence or hazard function, e.g. $a(t) = 1/(1-t)$. For small dt:

$$P(\theta \text{ diverges} \in (t, t+dt)) = \frac{a(t)dt}{m}$$

  where m is the number of previous points that have followed this path.
- If θ_i reaches a branch point between two paths, it picks a branch in proportion to the number of points that have followed that path.

Dirichlet Diffusion Trees (DFT)

Generating from a DFT:

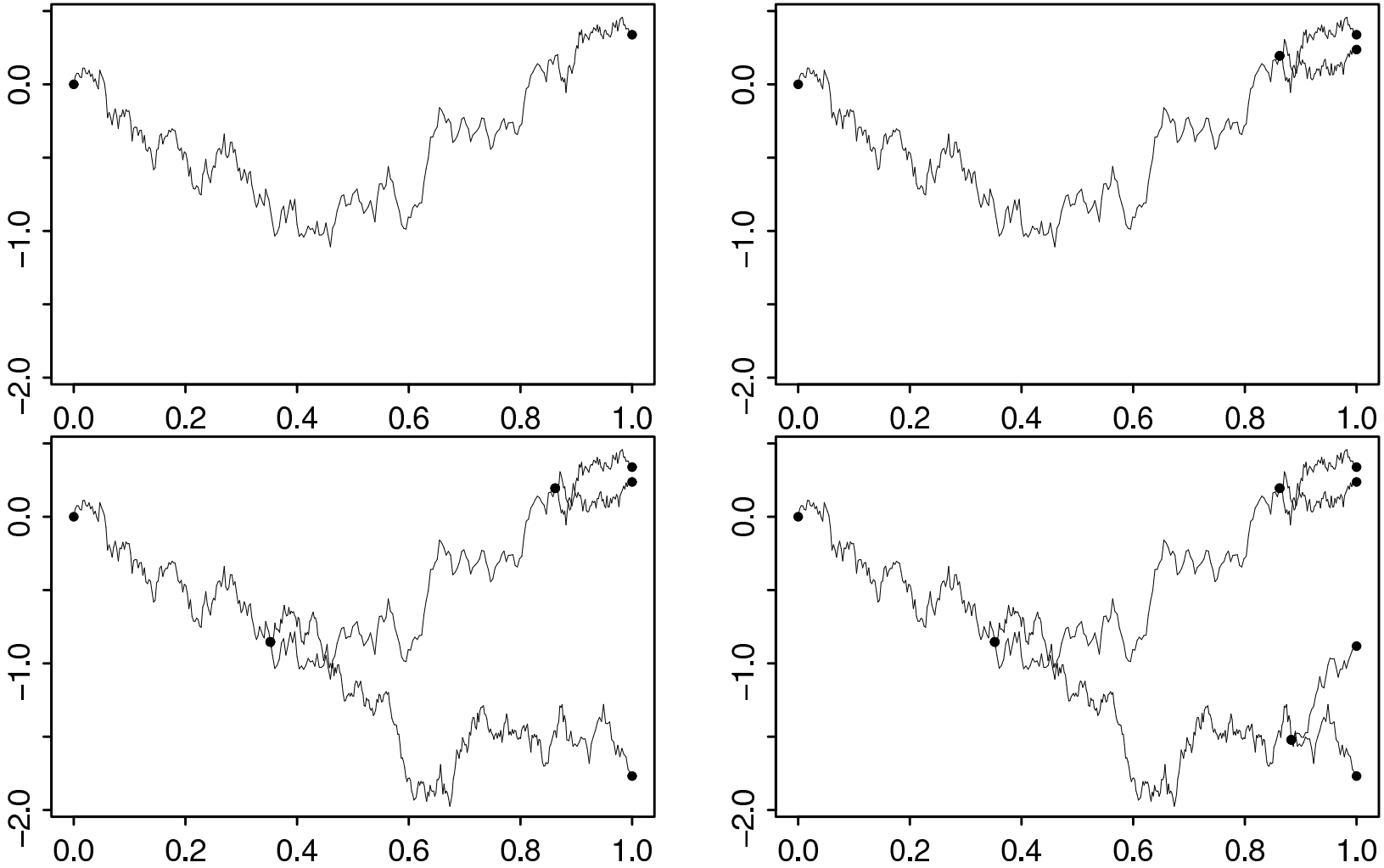


Figure from Neal, 2001.

Dirichlet Diffusion Trees (DFT)

Some samples from DFT priors:

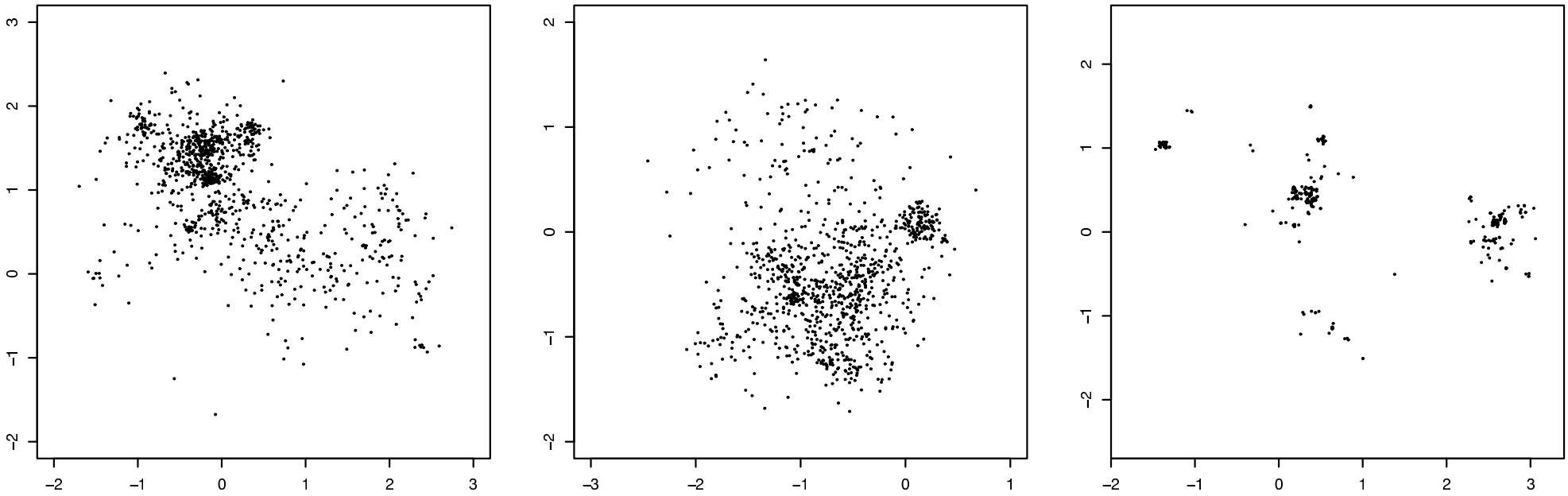


Figure from Neal, 2001.

Kingman's Coalescent

(Kingman, 1982)

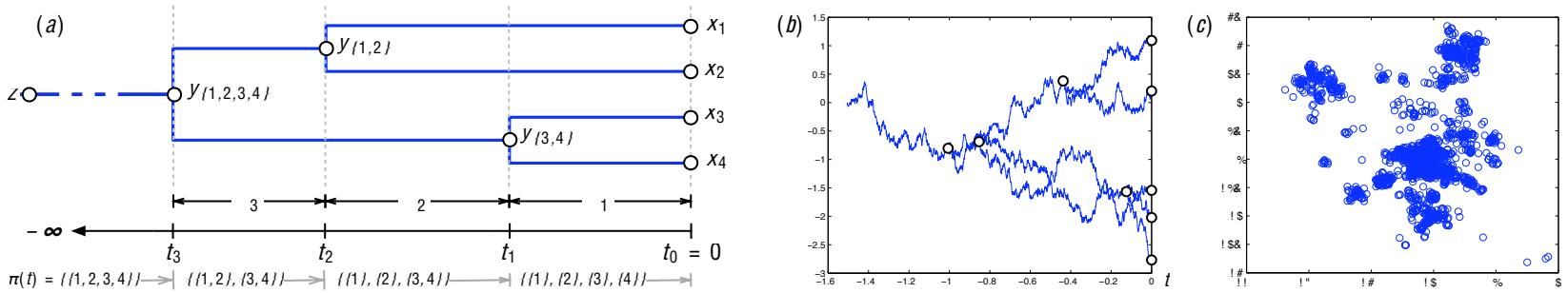
A model for genealogies. Working backwards in time from $t = 0$

- Generate tree with n individuals at leaves by merging lineages backwards in time.
- Every pair of lineages merges independently with rate 1.
- Time of first merge for n individuals is $t_1 \sim \mathrm{Exp}(n(n-1)/2)$, etc

This generates w.p.1. a binary tree where all lineages are merged at $t = -\infty$.

The coalescent results in the limit $n \to \infty$ and is infinitely exchangeable over individuals and has uniform marginals over tree topologies.

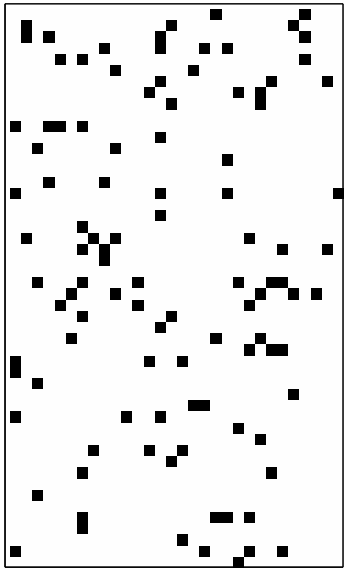
The coalescent has been used as the basis of a hierarchical clustering algorithm by (Teh, Daumé, Roy, 2007)



Latent Feature Models

Indian Buffet Processes: Distributions on Sparse Binary Matrices

- Rows are data points
- Columns are latent features
- There are **infinitely** many latent features
- Each data point can have *multiple* features



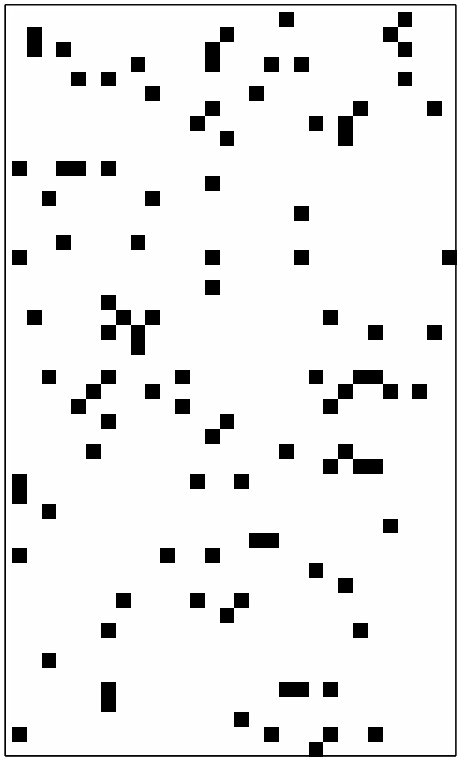
Another way of thinking about this:

- there are multiple overlapping clusters
- each data point can belong to several clusters simultaneously.

If there are K features, then there are 2^K possible binary latent representations for each data point.

(Griffiths and Ghahramani, 2005)

From finite to infinite binary matrices



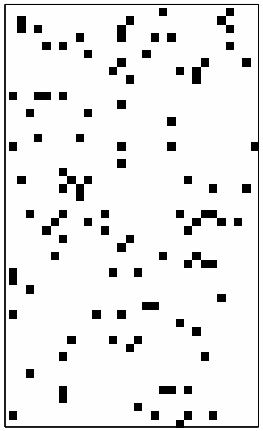
$z_{ik} = 1$ means object i has feature k:

$$z_{ik} \sim \text{Bernoulli}(\theta_k)$$

$$\theta_k \sim \text{Beta}(\alpha/K, 1)$$

- Note that $P(z_{ik} = 1|\alpha) = E(\theta_k) = \frac{\alpha/K}{\alpha/K+1}$, so as K grows larger the matrix gets sparser.
- So if $\mathbf{Z}$ is $N \times K$, the expected number of nonzero entries is $N\alpha/(1 + \alpha/K) < N\alpha$.
- Even in the $K \to \infty$ limit, the matrix is expected to have a finite number of non-zero entries.

From finite to infinite binary matrices



We can integrate out vector of Beta variables θ, leaving:

$$
\begin{aligned}
P(\mathbf{Z}|\alpha) &= \int P(\mathbf{Z}|\theta)P(\theta|\alpha)d\theta \\
&= \prod_k \frac{\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(\frac{\alpha}{K})} \frac{\Gamma(1 + \frac{\alpha}{K})}{\Gamma(N + 1 + \frac{\alpha}{K})}
\end{aligned}
$$

The conditional feature assignments are:

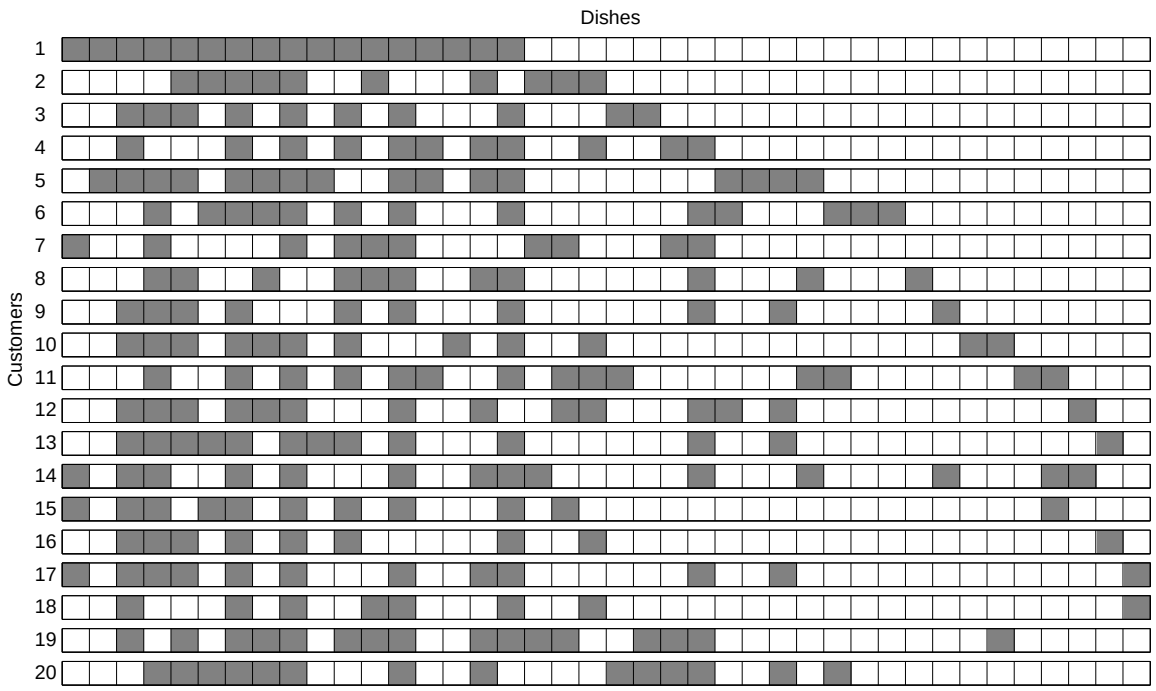
$$
P(z_{ik} = 1|\mathbf{z}_{-i,k}) = \int_0^1 P(z_{ik}|\theta_k)p(\theta_k|\mathbf{z}_{-i,k})\ d\theta_k = \frac{m_{-i,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}},
$$

where $\mathbf{z}_{-i,k}$ is the set of assignments of all objects, not including i, for feature k, and $m_{-i,k}$ is the number of objects having feature k, not including i.
We can take limit as $K \to \infty$.

"Rich get richer", like in Chinese Restaurant Processes.
Infinite vector of Beta random variables θ related to Beta process.

Indian buffet process

(Griffiths and Ghahramani, 2005)

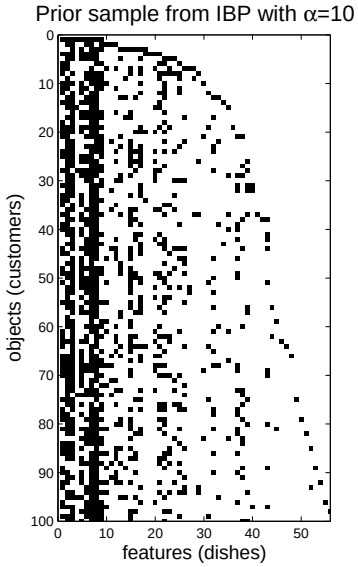


"Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes"

- First customer starts at the left of the buffet, and takes a serving from each dish, stopping after a Poisson(α) number of dishes.
- The nth customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself dish k with probability m_k/n, and trying a Poisson(α/n) number of new dishes.
- The customer-dish matrix is the feature matrix, $\mathbf{Z}$.

Properties of the Indian buffet process

$$P([\mathbf{Z}]|\alpha) = \exp\{-\alpha H_N\}\frac{\alpha^{K_+}}{\prod_{h>0} K_h!}\prod_{k\leq K_+}\frac{(N-m_k)!(m_k-1)!}{N!}$$



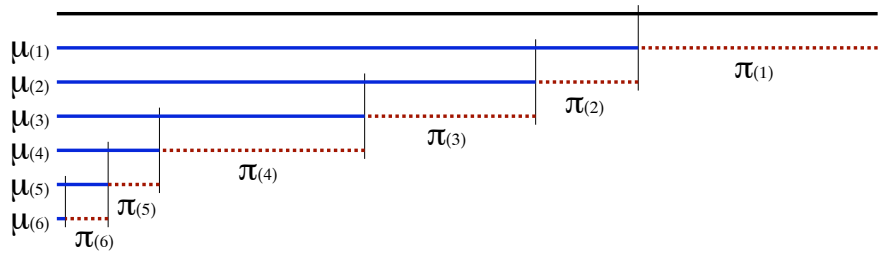


Figure 1: Stick-breaking construction for the DP and IBP. The black stick at top has length 1. At each iteration the vertical black line represents the break point. The brown dotted stick on the right is the weight obtained for the DP, while the blue stick on the left is the weight obtained for the IBP.

Shown in (Griffiths and Ghahramani, 2005):

- It is infinitely exchangeable.
- The number of ones in each row is Poisson(α)
- The expected total number of ones is αN.
- The number of nonzero columns grows as $O(\alpha \log N)$.

Additional properties:

- Has a stick-breaking representation (Teh, Görür, Ghahramani, 2007)
- Has as its de Finetti mixing distribution the Beta process (Thibaux and Jordan, 2007)

Completely Random Measures

(Kingman, 1967)

measurable space: Θ with σ-algebra Ω
measure: function $\mu : \Omega \to [0, \infty]$ assigning to each measurable set a non-neg. real
random measure: measures are drawn from some distribution on measures: $\mu \sim P$
completely random measure (CRM): the values that μ takes on disjoint subsets are independent: $\mu(A) \perp\!\!\!\perp \mu(B) \quad \text{if} \quad A \cap B = \emptyset$

CRMs can be represented as sum of nonrandom measure, atomic measure with fixed atoms but random masses, and atomic measure with random atoms and masses.

$$\mu = \mu_0 + \sum_{i=1}^{N \text{ or } \infty} u'_i \delta_{\phi'_i} + \sum_{i=1}^{M \text{ or } \infty} u_i \delta_{\phi_i}$$

We can write

$$\mu \sim \mathrm{CRM}(\mu_0, \Lambda, \{\phi'_i, F_i\})$$

where $u'_i \sim F_i$ and $\{u_i, \phi_i\}$ are drawn from a Poisson process on $(0, \infty] \times \Theta$ with rate measure Λ called the Lévy measure.

Examples:

Gamma process, Beta process (Hjort, 1990), Stable-beta process (Teh and Görür, *this NIPS*, 2009).

Beta Processes

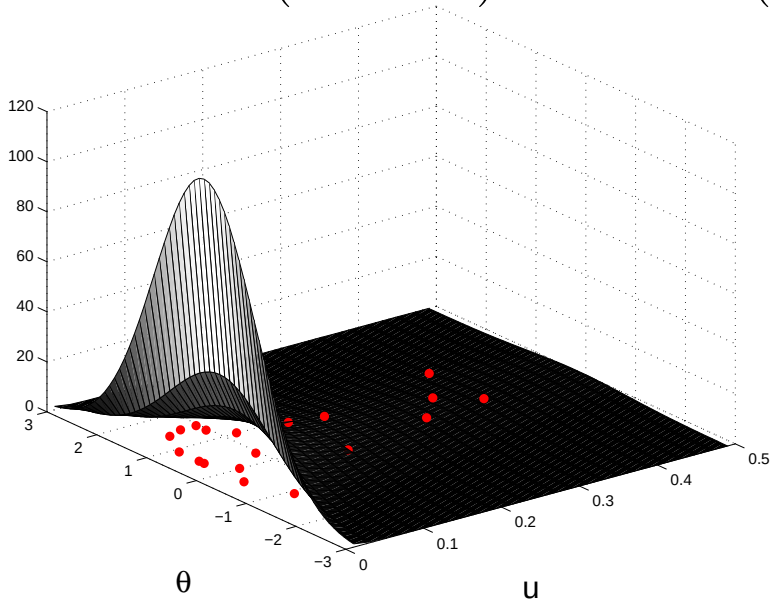
(Hjort, 1990)

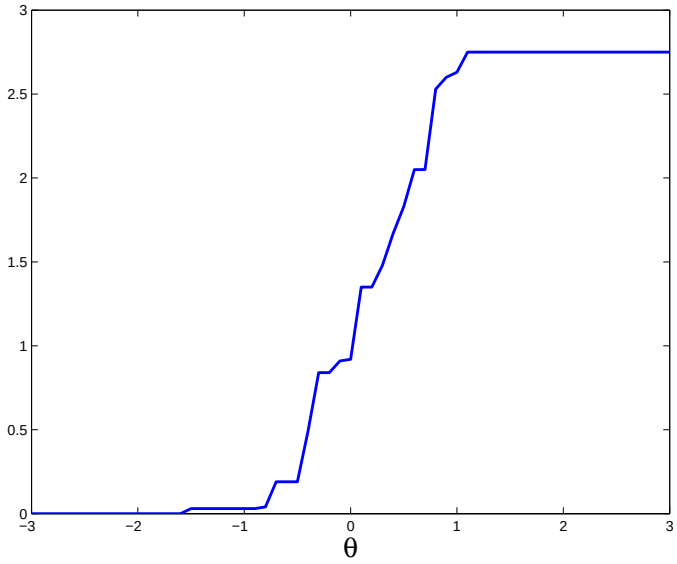
CRMs can be represented as sum of nonrandom measure, atomic measure with fixed atoms but random masses, and atomic measure with random atoms and masses.

$$\mu = \mu_0 + \sum_{i=1}^{N \text{ or } \infty} u_i' \delta_{\phi_i'} + \sum_{i=1}^{M \text{ or } \infty} u_i \delta_{\phi_i}$$

For a beta process we have $\mu \sim \mathrm{CRM}(0, \Lambda, \{\})$ where $\{u_i, \phi_i\}$ are drawn from a Poisson process on $(0, \infty] \times \Theta$ with rate Lévy measure:

$$\Lambda(du \times d\theta) = \alpha\, c\, u^{-1}(1-u)^{c-1} du\, H(d\theta)$$





A Few Topics I Didn't Cover

Models for time series

- infinite Hidden Markov Model / HDP-HMM (Beal, Ghahramani, Rasmussen, 2002; Teh, Jordan, Beal, Blei, 2006; Fox, Sudderth, Jordan, Willsky, 2008)
- infinite factorial HMM / Markov IBP (van Gael, Teh, Ghahramani, 2009)
- beta process HMM (Fox, Sudderth, Jordan, Willsky, 2009)

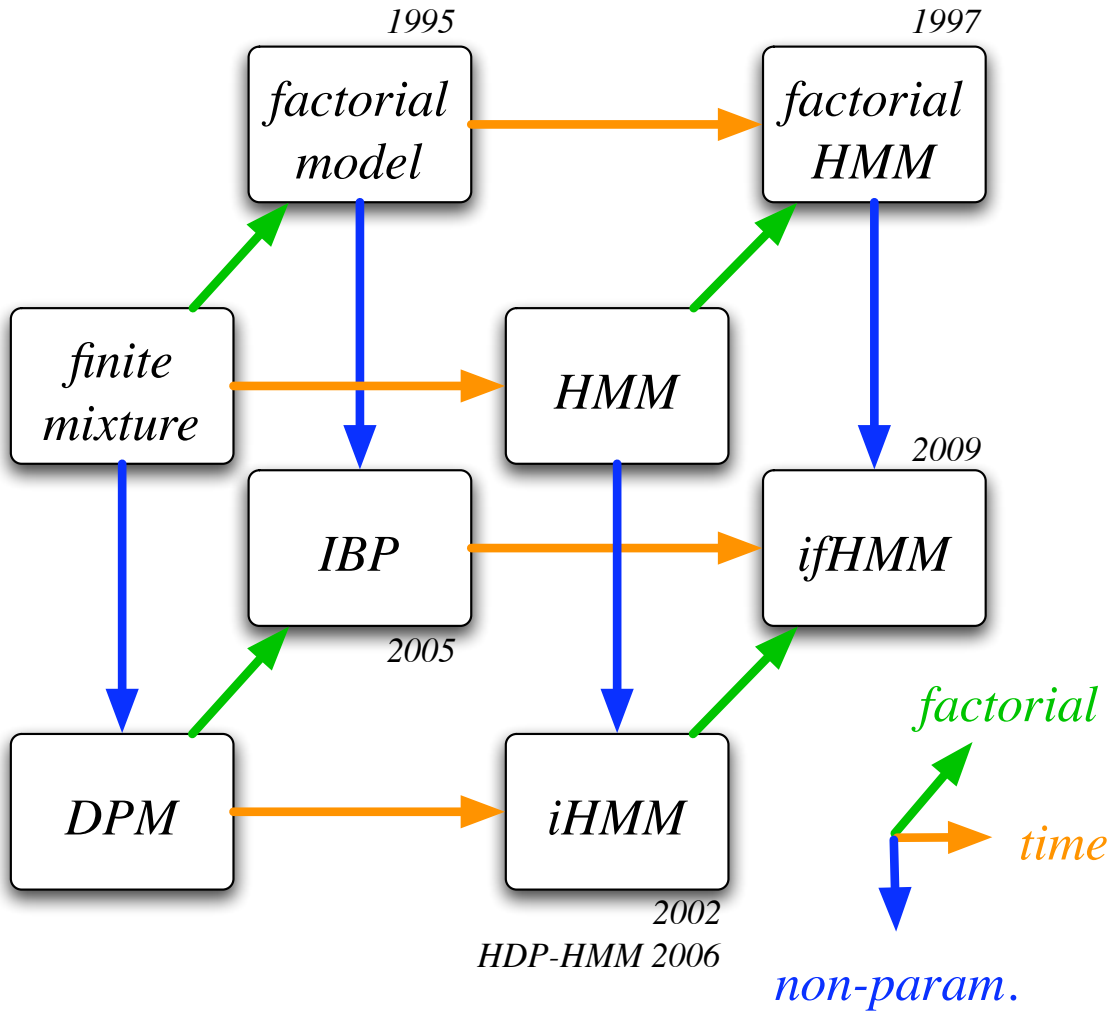
Hierarchical models for sharing structure

- hierarchical Dirichlet processes (Teh, Jordan, Beal, Blei, 2006)
- hierarchical beta processes (Thibaux, Jordan, 2007)

Many generalisations.

Inference!!

Some Relationships



Thanks for your patience!