Bayesian Nonparametric Modelling:
Dirichlet Processes, Hierarchical Dirichlet Processes, Indian Buffet Processes

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

April 17, April 25, 2008 / MLII

Outline

Bayesian Nonparametric Modelling
- Gaussian Processes
- De Finetti's Theorem
- Pólya Urn Scheme

Dirichlet Processes

Representations of Dirichlet Processes

Some Applications of Dirichlet Processes

Hierarchical Dirichlet Processes

Nested and Dependent Dirichlet Processes

Indian Buffet Processes

Modelling Data

All models are wrong, but some are useful.

—George E. P. Box, Norman R. Draper (1987).

- Models are never correct for real world data.
- How do we deal with model misfit?
  1. Model selection or averaging;
  2. Quantify closeness to true model, and optimality of fitted model;
  3. Increase the flexibility of your model class.

Nonparametric Modelling

- What is a nonparametric model?
  1. A parametric model where the number of parameters increases with data;
  2. A really large parametric model;
  3. A model over infinite dimensional function or measure spaces.
- Why nonparametric models in Bayesian theory of learning?
  1. broad class of priors that allows data to "speak for itself";
  2. side-step model selection and averaging.
- How do we deal with the infinite parameter space?
  1. Marginalize out all but a finite number of parameters;
  2. Define infinite space implicitly (akin to the kernel trick) using either Kolmogorov Consistency Theorem or de Finetti's theorem.

Gaussian Processes

A *Gaussian process* (GP) is a random function $f : \mathbb{X} \to \mathbb{R}$ such that for any finite set of input points $x_1, \dots, x_n$,

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} c(x_1, x_1) & \dots & c(x_1, x_n) \\ \vdots & \ddots & \vdots \\ c(x_n, x_1) & \dots & c(x_n, x_n) \end{bmatrix} \right)$$

where the parameters are the mean function $m(x)$ and covariance kernel $c(x, y)$.

- The above finite dimensional marginal distributions are *consistent*, which guarantees existence of GPs via the *Kolmogorov Consistency Theorem*.
- GPs can be visualized by iterative sampling $f(x_n)|f(x_1), \dots, f(x_{n-1})$ on a sequence of input points $x_1, x_2, \dots$.

[Rasmussen and Williams 2006]

De Finetti's Theorem

Let $\theta_1, \theta_2, \ldots$ be an infinite sequence of random variables with joint distribution p. If for all $n \geq 1$, and all permutations $\sigma \in \Sigma_n$ on n objects,

$$p(\theta_1, \ldots, \theta_n) = p(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)})$$

That is, the sequence is infinitely exchangeable. Then there exists a latent random parameter G such that:

$$p(\theta_1, \ldots, \theta_n) = \int \rho(G) \prod_{i=1}^{n} \rho(\theta_i | G) dG$$

where ρ is a joint distribution over G and θ_i's.

- θ_i's are *independent* given G.
- Sufficient to define p through the conditionals $p(\theta_n | \theta_1, \ldots, \theta_{n-1})$.
- G can be *infinite dimensional* (indeed it is often a *random measure*).
- The set of infinitely exchangeable sequences is *convex* and it is an important theoretical topic to study the set of *extremal points*.
- Partial exchangeability: Markov, arrays...

Pólya Urn Scheme

Let $\alpha \geq 0$ and H be some distribution. The Pólya urn scheme operates as follows:

1. *Draw $\theta_1 \sim H$.*
2. *For $n = 2, 3, \ldots$, let*

$$\theta_n | \theta_1, \ldots, \theta_{n-1} \sim \frac{1}{n-1+\alpha} \sum_{i=1}^{n-1} \delta_{\theta_i} + \frac{\alpha}{n-1+\alpha} H$$

   where δ_θ is a point mass at θ.

That is, with probability $\frac{1}{n-1+\alpha}$, $\theta_n = \theta_i$, while with probability $\frac{\alpha}{n-1+\alpha}$ we have that θ_n is drawn from H.

- The Pólya urn scheme generates a sequence $\theta_1, \theta_2, \ldots$
- It is infinitely exchangeable.
- Also known as Blackwell-MacQueen urn scheme.

[Blackwell and MacQueen 1973]

Pólya Urn Scheme

Proof of exchangeability:

Suppose H is non-atomic.

Let $\theta_1, \ldots, \theta_K^*$ be the unique values, and $m_{nk} = \sum_{i=1}^{n} \mathbf{1}(\theta_i = \theta_k^*)$. Then by collecting terms in the generative process probabilities:

$$p(\theta_1, \ldots, \theta_n) = \frac{\alpha^K \prod_{k=1}^{K} h(\theta_k^*)(m_{nk} - 1)!}{\prod_{i=1}^{n} i - 1 + \alpha}$$

where $h(\theta)$ is density of θ under H.

- If H has atoms, above proof works too, but we need to define the *clustering structure* more carefully.
- It is possible to define a sequence of joint probabilities $p_n(\theta_1, \ldots, \theta_n)$ for $n \geq 1$, such that each p_n is *finitely exchangeable* but not infinitely exchangeable. We also need *consistency*:

$$\int p_{n+1}(\theta_1, \ldots, \theta_{n+1}) d\theta_{n+1} = p_n(\theta_1, \ldots, \theta_n)$$

- What is the de Finetti measure of the Pólya urn scheme?

Outline

Bayesian Nonparametric Modelling

Dirichlet Processes
- Dirichlet Distributions
- Definition
- Parameters of Dirichlet Processes

Representations of Dirichlet Processes

Some Applications of Dirichlet Processes

Hierarchical Dirichlet Processes

Nested and Dependent Dirichlet Processes

Indian Buffet Processes

A Very Little Measure Theory

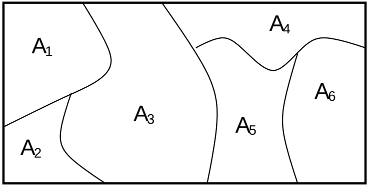
- A σ-algebra Σ is a family of subsets of a set Θ such that
  - Σ is not empty;
  - If $A \in \Sigma$ then $\Theta \backslash A \in \Sigma$;
  - If $A_1, A_2, \ldots \in \Sigma$ then $\cup_{i=1}^{\infty} A_i \in \Sigma$.
- (Θ, Σ) is a measure space and $A \in \Sigma$ are the measurable sets.
- A measure μ over (Θ, Σ) is a function $\mu : \Sigma \to [0, \infty]$ such that
  - $\mu(\emptyset) = 0$;
  - If $A_1, A_2, \ldots \in \Sigma$ are disjoint then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.
  - Everything we consider here will be measurable.
  - A probability measure is one where $\mu(\Theta) = 1$.
  - We will identify probability measures as equivalent to distributions over random variables X taking on values in Θ. Basically $p(X \in A) = \mu(A)$ for an event $A \in \Sigma$.

Dirichlet Processes

A *Dirichlet Process* (DP) is a random probability measure G over (Θ, Σ) such that for any finite set of partitions $A_1 \dot{\cup} \dots \dot{\cup} A_K = \Theta$, the random vector

$$(G(A_1), \dots, G(A_K))$$

is Dirichlet distributed.



- Reminder: probability measures are functions, and above definition is very similar to that of Gaussian processes.
- Kolmogorov Consistency Theorem can be applied again to show that random functions $G : \Sigma \to [0, 1]$ exists, but there are technical difficulties.

[Ferguson 1973]

Dirichlet Distributions

- A *Dirichlet distribution* is a distribution over the K-dimensional probability simplex:

$$\Delta_K = \left\{(\pi_1, \ldots, \pi_K) \,:\, \pi_k \geq 0, \textstyle\sum_k \pi_k = 1\right\}$$

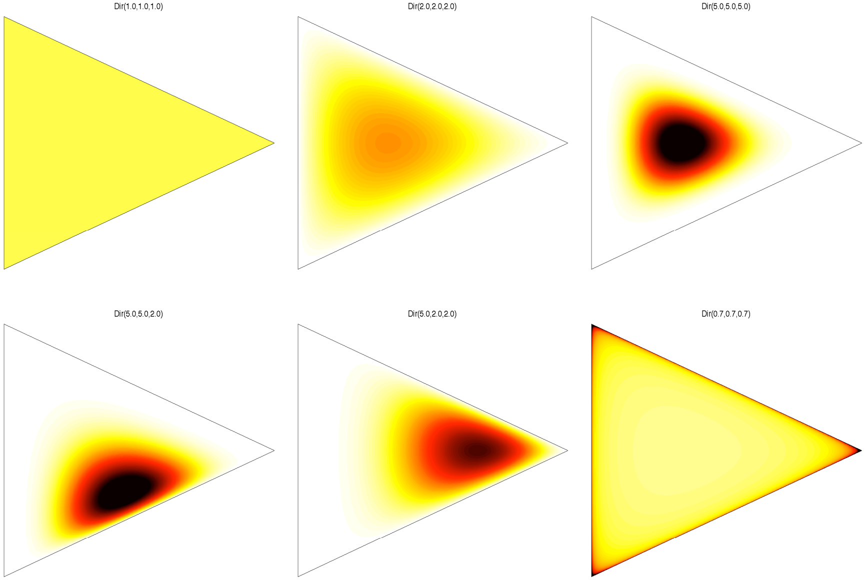
- We say $(\pi_1, \ldots, \pi_K)$ is Dirichlet distributed,

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

with parameters $(\alpha_1, \ldots, \alpha_K)$, if

$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{n} \pi_k^{\alpha_k - 1}$$

Dirichlet Distributions



Dirichlet Distributions: Agglomerative Property

- Combining entries of probability vectors preserves Dirichlet property, for example:

$$
\begin{aligned}
& (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\
\Rightarrow \quad & (\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)
\end{aligned}
$$

- Generally, if $(I_1, \ldots, I_j)$ is a partition of $(1, \ldots, n)$:

$$\left(\sum_{i \in I_1} \pi_i, \ldots, \sum_{i \in I_j} \pi_i \right) \sim \text{Dirichlet} \left(\sum_{i \in I_1} \alpha_i, \ldots, \sum_{i \in I_j} \alpha_i \right)$$

Dirichlet Distributions: Decimative Property

- The converse of the agglomerative property is also true, for example if:

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\
(\tau_1, \tau_2) &\sim \text{Dirichlet}(\alpha_1 \beta_1, \alpha_1 \beta_2)
\end{aligned}
$$

with $\beta_1 + \beta_2 = 1$,

$$
\Rightarrow \quad (\pi_1 \tau_1, \pi_1 \tau_2, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 \beta_1, \alpha_2 \beta_2, \alpha_2, \ldots, \alpha_K)
$$

Dirichlet Processes

- A Dirichlet process (DP) is an "infinitely decimated" Dirichlet variable:

$$
\begin{aligned}
1 &\sim \text{Dirichlet}(\alpha) \\
(\pi_1, \pi_2) &\sim \text{Dirichlet}(\alpha/2, \alpha/2) && \pi_1 + \pi_2 = 1 \\
(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) &\sim \text{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4) && \pi_{i1} + \pi_{i2} = \pi_i \\
&\vdots
\end{aligned}
$$

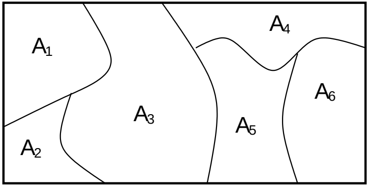
- Each decimation step involves drawing from a Beta distribution (a Dirichlet with 2 components) and multiplying into the relevant entry.
- Demo: DPgenerate

Dirichlet Processes

A *Dirichlet Process* (DP) is a random probability measure G over (Θ, Σ) such that for any finite set of partitions $A_1 \dot\cup \ldots \dot\cup A_K = \Theta$, the random vector

$$(G(A_1), \ldots, G(A_K))$$

is Dirichlet distributed.



- Reminder: probability measures are functions, and above definition is very similar to that of Gaussian processes.
- Kolmogorov Consistency Theorem can be applied again to show that random functions $G : \Sigma \to [0, 1]$ exists, but there are technical difficulties.

[Ferguson 1973]

Parameters of Dirichlet Processes

- A DP has two parameters:
  - *Base distribution* H, which is like the *mean* of the DP.
  - *Strength parameter* α, which is like an *inverse-variance* of the DP.
- We write:

$$G \sim \mathrm{DP}(\alpha, H)$$

if for any partition $(A_1, \ldots, A_K)$ of Θ:

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

- The first two cumulants of the DP:

$$\text{Expectation:} \qquad \mathbb{E}[G(A)] = H(A)$$

$$\text{Variance:} \qquad \mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

where A is any measurable subset of Θ.

Outline

Bayesian Nonparametric Modelling

Dirichlet Processes

Representations of Dirichlet Processes

- Posterior Dirichlet Processes
- Pólya Urn Scheme
- Chinese Restaurant Process
- Stick-breaking Construction
- Extensions of Dirichlet Processes

Some Applications of Dirichlet Processes

Hierarchical Dirichlet Processes

Nested and Dependent Dirichlet Processes

Indian Buffet Processes

Representations of Dirichlet Processes

- Suppose $G \sim \mathrm{DP}(\alpha, H)$. G is a (random) probability measure over Θ. We can treat it as a distribution over Θ. Let

$$\theta_1, \ldots, \theta_n \sim G$$

  be random variables with distribution G.

- We saw in the demo that draws from a Dirichlet process seem to be discrete distributions. If so, then:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

  and there is positive probability that sets of θ_i's can take on the same value θ_k^* for some k, i.e. the θ_i's cluster together.

- We are concerned with representations of Dirichlet processes based upon both the clustering property and the sum of point masses.

Posterior Dirichlet Processes

- Suppose G is DP distributed, and θ is G distributed:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned}$$

- This gives $p(G)$ and $p(\theta|G)$.
- We are interested in:

$$\begin{aligned} p(\theta) &= \int p(\theta|G) p(G)\, dG \\ p(G|\theta) &= \frac{p(\theta|G)p(G)}{p(\theta)} \end{aligned}$$

Posterior Dirichlet Processes

Conjugacy between Dirichlet Distribution and Multinomial.

- Consider:

$$(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$$
$$z | (\pi_1, \dots, \pi_K) \sim \text{Discrete}(\pi_1, \dots, \pi_K)$$

  z is a multinomial variate, taking on value $i \in \{1, \dots, n\}$ with probability π_i.

- Then:

$$z \sim \text{Discrete}\left(\frac{\alpha_1}{\sum_i \alpha_i}, \dots, \frac{\alpha_K}{\sum_i \alpha_i}\right)$$
$$(\pi_1, \dots, \pi_K) | z \sim \text{Dirichlet}(\alpha_1 + \delta_1(z), \dots, \alpha_K + \delta_K(z))$$

  where $\delta_i(z) = 1$ if z takes on value i, 0 otherwise.

- Converse also true.

Posterior Dirichlet Processes

- Fix a partition $(A_1, \ldots, A_K)$ of Θ. Then

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$
$$P(\theta \in A_i | G) = G(A_i)$$

- Using Dirichlet-multinomial conjugacy,

$$P(\theta \in A_i) = H(A_i)$$
$$(G(A_1), \ldots, G(A_K))|\theta \sim \mathrm{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \ldots, \alpha H(A_K) + \delta_\theta(A_K))$$

- The above is true for every finite partition of Θ. In particular, taking a really fine partition,

$$p(d\theta) = H(d\theta)$$

- Also, the posterior $G|\theta$ is also a Dirichlet process:

$$G|\theta \sim \mathrm{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$$

Posterior Dirichlet Processes

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right) \end{aligned}$$

Pólya Urn Scheme

- First sample:

$$\theta_1|G \sim G \qquad G \sim \mathrm{DP}(\alpha, H)$$

$$\iff \qquad \theta_1 \sim H \qquad G|\theta_1 \sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1})$$

- Second sample:

$$\theta_2|\theta_1, G \sim G \qquad G|\theta_1 \sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1})$$

$$\iff \qquad \theta_2|\theta_1 \sim \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1} \qquad G|\theta_1, \theta_2 \sim \mathrm{DP}(\alpha+2, \tfrac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha+2})$$

- n^{th} sample

$$\theta_n|\theta_{1:n-1}, G \sim G \qquad G|\theta_{1:n-1} \sim \mathrm{DP}(\alpha+n-1, \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1})$$

$$\iff \qquad \theta_n|\theta_{1:n-1} \sim \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} \qquad G|\theta_{1:n} \sim \mathrm{DP}(\alpha+n, \tfrac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n})$$

Pólya Urn Scheme

- Pólya urn scheme produces a sequence $\theta_1, \theta_2, \ldots$ with the following conditionals:

$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- Picking balls of different colors from an urn:
  - Start with no balls in the urn.
  - with probability $\propto \alpha$, draw $\theta_n \sim H$, and add a ball of that color into the urn.
  - With probability $\propto n - 1$, pick a ball at random from the urn, record θ_n to be its color, return the ball into the urn and place a second ball of same color into urn.
- Pólya urn scheme is like a “representer” for the DP—a finite projection of an infinite object G.

Exchangeability and De Finetti's Theorem

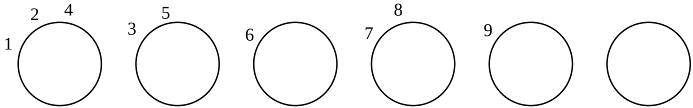
- Starting with a DP, we constructed the Pólya urn scheme.
- De Finetti's Theorem gives the converse.
- Since θ_i are iid G, their joint distribution is invariant to permutations, thus $\theta_1, \theta_2, \ldots$ are infinitely exchangeable.
- Thus a random measures must exist making them iid.
- This is G.

Chinese Restaurant Process

- Draw $\theta_1, \ldots, \theta_n$ from a Pólya urn scheme.
- They take on $K < n$ distinct values, say $\theta_1^*, \ldots, \theta_K^*$.
- This defines a partition of $1, \ldots, n$ into K clusters, such that if i is in cluster k, then $\theta_i = \theta_k^*$.
- Random draws $\theta_1, \ldots, \theta_n$ from a Pólya urn scheme induces a random partition of $1, \ldots, n$.
- The induced distribution over partitions is a *Chinese restaurant process* (CRP).

Chinese Restaurant Process

- Generating from the CRP:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
  - Customers $\Leftrightarrow$ integers, tables $\Leftrightarrow$ clusters.
- The CRP exhibits the *clustering property* of the DP.
- Rich-gets-richer effect implies small number of large clusters.
- Expected number of clusters is $K = O(\alpha \log n)$.



Chinese Restaurant Process

- To get back from the CRP to Pólya urn scheme, simply draw

$$\theta_k^* \sim H$$

for $k = 1, \ldots, K$, then for $i = 1, \ldots, n$ set

$$\theta_i = \theta_{z_i}^*$$

where z_i is the table that customer i sat at.

- The clustering (partition) is independent from the values assigned to each cluster.
- The CRP teases apart the clustering property of the DP, from the base distribution.

Stick-breaking Construction

- Returning to the posterior process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta|G &\sim G \end{aligned} \qquad \Leftrightarrow \qquad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H+\delta_\theta}{\alpha+1}) \end{aligned}$$

- Consider a partition $(\theta, \Theta\backslash\theta)$ of Θ. We have:

$$\begin{aligned} (G(\theta), G(\Theta\backslash\theta))|\theta &\sim \mathrm{Dirichlet}((\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\theta), (\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\Theta\backslash\theta)) \\ &= \mathrm{Dirichlet}(1, \alpha) \end{aligned}$$

- G has a point mass located at θ:

$$G = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \mathrm{Beta}(1, \alpha)$$

  and G' is the (renormalized) probability measure with the point mass removed.

- What is G'?

Stick-breaking Construction

- Currently, we have:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H) \\
\theta &\sim G
\end{aligned}
\qquad \Rightarrow \qquad
\begin{aligned}
\theta &\sim H \\
G|\theta &\sim \mathrm{DP}(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1}) \\
G &= \beta \delta_\theta + (1 - \beta) G' \\
\beta &\sim \mathrm{Beta}(1, \alpha)
\end{aligned}
$$

- Consider a further partition $(\theta, A_1, \ldots, A_K)$ of Θ:

$$
\begin{aligned}
&(G(\theta), G(A_1), \ldots, G(A_K)) \\
=&(\beta, (1 - \beta) G'(A_1), \ldots, (1 - \beta) G'(A_K)) \\
\sim&\, \mathrm{Dirichlet}(1, \alpha H(A_1), \ldots, \alpha H(A_K))
\end{aligned}
$$

- The agglomerative/decimative property of Dirichlet implies:

$$
\begin{aligned}
(G'(A_1), \ldots, G'(A_K))|\theta &\sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)) \\
G' &\sim \mathrm{DP}(\alpha, H)
\end{aligned}
$$

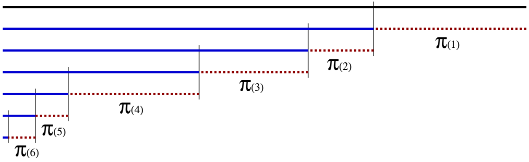
Stick-breaking Construction

- We have:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H)\\
G &= \beta_1\delta_{\theta_1^*} + (1-\beta_1)G_1\\
G &= \beta_1\delta_{\theta_1^*} + (1-\beta_1)(\beta_2\delta_{\theta_2^*} + (1-\beta_2)G_2)\\
&\vdots\\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}
\end{aligned}
$$

where

$$\pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i) \qquad \beta_k \sim \mathrm{Beta}(1,\alpha) \qquad \theta_k^* \sim H$$



Stick-breaking Construction

- We call the construction for π_1, π_2 the *stick-breaking construction*.
- Also known as the GEM distribution, write $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$.
- Starting with a DP, we showed that draws from the DP looks like a sum of point masses, with masses drawn from a stick-breaking construction.
- The steps are limited by assumptions of regularity on Θ and smoothness on H.
- [Sethuraman 1994] started with the stick-breaking construction, and showed that draws are indeed DP distributed, under very general conditions.

Representations of Dirichlet Processes

- Posterior Dirichlet process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned} \qquad \Longleftrightarrow \qquad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}\left(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1}\right) \end{aligned}$$

- Pólya urn scheme:

$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- Chinese restaurant process:

$$p(\text{customer } n \text{ sat at table } k | \text{past}) = \begin{cases} \frac{n_k}{n-1+\alpha} & \text{if occupied table} \\ \frac{\alpha}{n-1+\alpha} & \text{if new table} \end{cases}$$

- Stick-breaking construction:

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \qquad \beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \theta_k^* \sim H \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

Extensions of Dirichlet Processes

- Two-parameter generalization of the Chinese restaurant process:

$$p(\text{customer } n \text{ sat at table } k|\text{past}) = \begin{cases} \frac{n_k - d}{n-1+\alpha} & \text{if occupied table} \\ \frac{\alpha+dK}{n-1+\alpha} & \text{if new table} \end{cases}$$

  Gives the *Pitman-Yor process*.

- Other stick-breaking constructions:

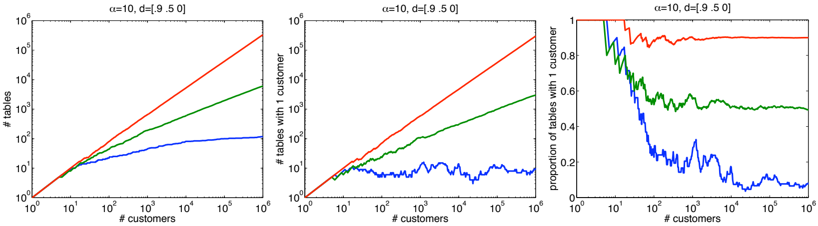
$$\pi_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i) \qquad \beta_k \sim \text{Beta}(a_k, b_k) \qquad \theta_k^* \sim H \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

  $a_k = 1 - d$, $b_k = \alpha + dk$ gives Pitman-Yor process.

[Pitman and Yor 1997, Perman et al. 1992]

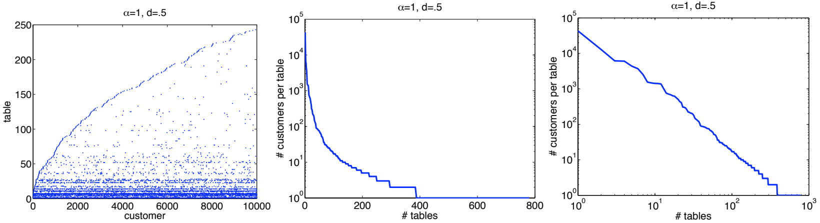
Pitman-Yor Processes

- Two salient features of the Pitman-Yor process:
  - With more occupied tables, the chance of even more tables becomes higher.
  - Tables with smaller occupancy numbers tend to have lower chance of getting new customers.
- The above means that Pitman-Yor processes produce Zipf's Law type behaviour, with $K = O(\alpha n^d)$.

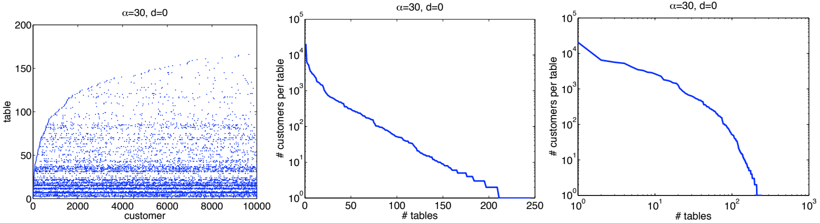


Pitman-Yor Processes

Draw from a Pitman-Yor process



Draw from a Dirichlet process



Normalized Gamma Processes

- A *gamma distribution* is a distribution over $[0, \infty)$. A gamma distributed variable $\gamma \sim \text{Gamma}(a, b)$ has density:

$$p(\gamma) = \frac{b^a}{\Gamma(a)} \gamma^{a-1} e^{-b\gamma}$$

- We can construct a Dirichlet variable by normalizing gamma variables:

$$\begin{aligned} \gamma_k &\sim \text{Gamma}(\alpha_k, 1) \\ (\pi_1, \ldots, \pi_K) &= \tfrac{1}{\sum_{k=1}^K \gamma_k} (\gamma_1, \ldots, \gamma_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \end{aligned}$$

- Similarly a DP can be constructed by normalizing a *gamma process*.

Normalized Gamma Processes

- A gamma process $\tilde{G} \sim \Gamma\mathrm{P}(\tilde{H})$ is a random measure satisfying:
  - $\tilde{G}(A) \sim \mathrm{Gamma}(\tilde{H}(A))$ for $A \in \Sigma$;
  - $\tilde{G}(A)$, $\tilde{G}(B)$ independent if $A \cup B = \emptyset$.
- A gamma process is a *completely random measure*—a random measure with independence on disjoint sets.
- This provides an avenue to generalize the DP by normalizing other completely random measures (e.g. *normalized generalized inverse Gaussian process*, *normalized stable process*).
- Another important example of completely random measures is the *beta process*.
- Completely random measures are strongly related to *Lévy processes*, which are in turn strongly related to *infinitely divisible distributions*.

Outline

Bayesian Nonparametric Modelling

Dirichlet Processes

Representations of Dirichlet Processes

Some Applications of Dirichlet Processes

- Density Estimation
- Clustering
- Semiparametric Modelling
- Model Selection/Averaging

Hierarchical Dirichlet Processes

Nested and Dependent Dirichlet Processes

Indian Buffet Processes

Density Estimation

- Parametric density estimation (e.g. Gaussian, mixture models)
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Model: $x_i | w \sim F(\cdot | w)$
- Prior over parameters

$$p(w)$$

- Posterior over parameters

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|w)p(w|\mathbf{x})\, dw$$

Density Estimation

- Bayesian nonparametric density estimation with Dirichlet processes
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Model: $x_i \sim G$
- Prior over distributions

$$G \sim \mathrm{DP}(\alpha, H)$$

- Posterior over distributions

$$p(G|\mathbf{x}) = \frac{p(G)p(\mathbf{x}|G)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|G)p(G|\mathbf{x})\, dF = \int G(x_\star)p(G|\mathbf{x})\, dG$$

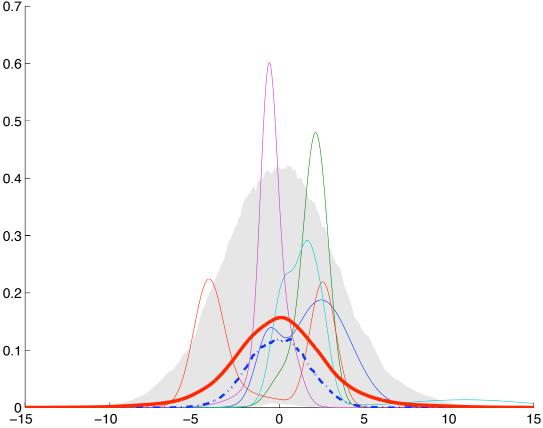
- *Not quite feasible, since G is a discrete distribution, in particular it has no density.*

Density Estimation

- Solution: Convolve the DP with a smooth distribution:

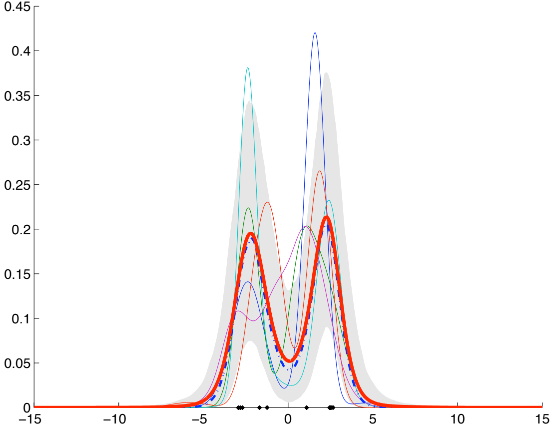
$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H) \\
F(\cdot) &= \int F(\cdot|\theta)dG(\theta) \\
x_i &\sim F_x
\end{aligned}
\qquad \Rightarrow \qquad
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \\
F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\
x_i &\sim F_x
\end{aligned}
$$

Density Estimation



$F(\cdot|\mu, \Sigma)$ is Gaussian with mean μ, covariance Σ.
$H(\mu, \Sigma)$ is Gaussian-inverse-Wishart conjugate prior.
Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws. Black: data points.

Density Estimation



$F(\cdot|\mu, \Sigma)$ is Gaussian with mean μ, covariance Σ.
$H(\mu, \Sigma)$ is Gaussian-inverse-Wishart conjugate prior.
Red: mean density. Blue: median density. Grey: 5-95 quantile. Others: draws. Black: data points.

Clustering

- Recall our approach to density estimation:

$$
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \sim \mathrm{DP}(\alpha, H) \\
F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\
x_i &\sim F_x
\end{aligned}
$$

- Above model equivalent to:

$$
\begin{aligned}
z_i &\sim \mathrm{Discrete}(\boldsymbol{\pi}) \\
\theta_i &= \theta_{z_i}^* \\
x_i|z_i &\sim F(\cdot|\theta_i) = F(\cdot|\theta_{z_i}^*)
\end{aligned}
$$

- This is simply a mixture model with an *infinite* number of components. This is called a *DP mixture model*.

Clustering

- DP mixture models are used in a variety of clustering applications, where the number of clusters is not known a priori.
- They are also used in applications in which we believe the number of clusters grows without bound as the amount of data grows.
- DPs have also found uses in applications beyond clustering, where the number of latent objects is not known or unbounded.
  - Nonparametric probabilistic context free grammars.
  - Visual scene analysis.
  - Infinite hidden Markov models/trees.
  - Haplotype inference.
  - ...
- In many such applications it is important to be able to model the same set of objects in different contexts.
- This corresponds to the problem of *grouped clustering* and can be tackled using *hierarchical Dirichlet processes*.

[Teh et al. 2006]

Semiparametric Modelling

- Linear regression model for inferring effectiveness of new medical treatments.

  $$y_{ij} = \beta^\top x_{ij} + b_i^\top z_{ij} + \epsilon_{ij}$$

  y_{ij} is outcome of jth trial on ith subject.

  x_{ij}, z_{ij} are predictors (treatment, dosage, age, health...).

  β are fixed-effects coefficients.

  b_i are random-effects subject-specific coefficients.

  ϵ_{ij} are noise terms.
- Care about inferring β. If x_{ij} is treatment, we want to determine $p(\beta > 0|\mathbf{x}, \mathbf{y})$.

Semiparametric Modelling

$$y_{ij} = \beta^\top x_{ij} + b_i^\top z_{ij} + \epsilon_{ij}$$

- Usually we assume Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Is this a sensible prior? Over-dispersion, skewness,...
- May be better to model noise nonparametrically,

$$\begin{aligned} \epsilon_{ij} &\sim F \\ F &\sim \mathrm{DP} \end{aligned}$$

- Also possible to model subject-specific random effects nonparametrically,

$$\begin{aligned} b_i &\sim G \\ G &\sim \mathrm{DP} \end{aligned}$$

Model Selection/Averaging

- Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Models: $p(\theta_k|M_k)$, $p(\mathbf{x}|\theta_k, M_k)$
- Marginal likelihood

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k)\, d\theta_k$$

- Model selection

$$M = \underset{M_k}{\operatorname{argmax}}\, p(\mathbf{x}|M_k)$$

- Model averaging

$$p(x_\star|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)p(M_k|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)\frac{p(\mathbf{x}|M_k)p(M_k)}{p(\mathbf{x})}$$

- *But: is this computationally feasible?*

Model Selection/Averaging

- Marginal likelihood is usually extremely hard to compute.

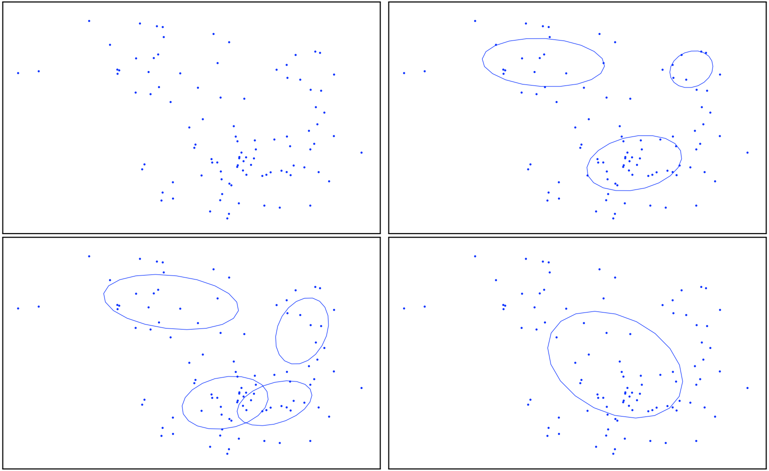
$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k)\, d\theta_k$$

- Model selection/averaging is to prevent underfitting and overfitting.
- But reasonable and proper Bayesian methods should not overfit [Rasmussen and Ghahramani 2001].
- Use a really large model M_∞ instead, and *let the data speak for themselves*.

Model Selection/Averaging

Clustering

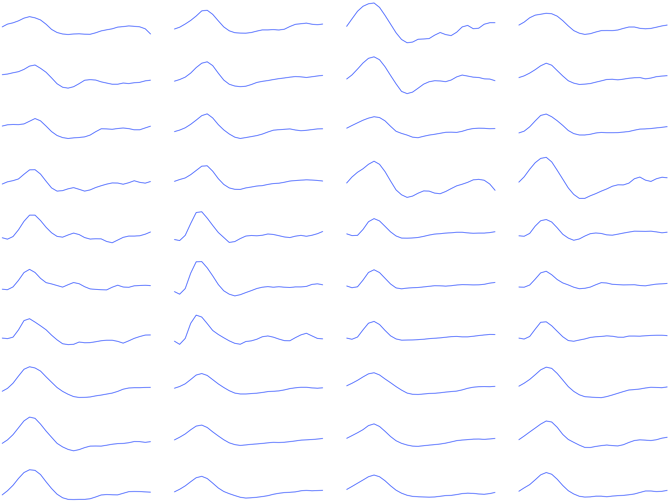
How many clusters are there?



Model Selection/Averaging

Spike Sorting

How many neurons are there?



[Görür 2007, Wood et al. 2006a]

Model Selection/Averaging

Topic Modelling

How many topics are there?

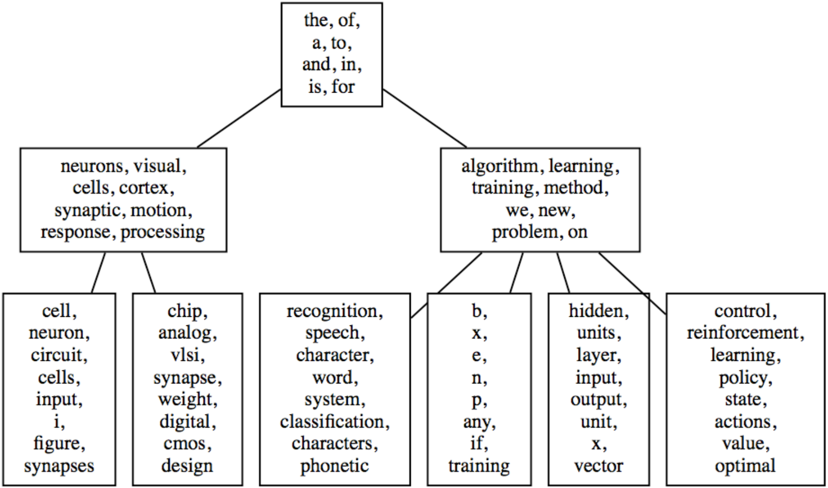


Figure from Blei et al. [Blei et al. 2004, Teh et al. 2006]

Model Selection/Averaging

Grammar Induction

How many grammar symbols are there?

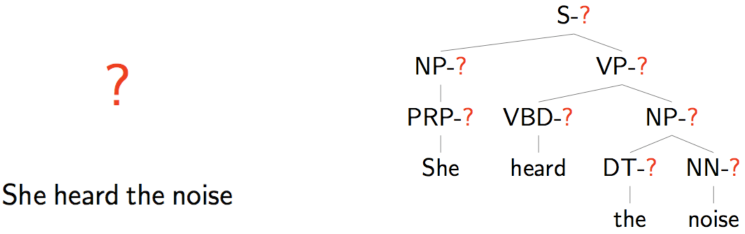


Figure from Liang et al. [Liang et al. 2007b, Finkel et al. 2007]

Model Selection/Averaging

Visual Scene Analysis

How many objects, parts, features?

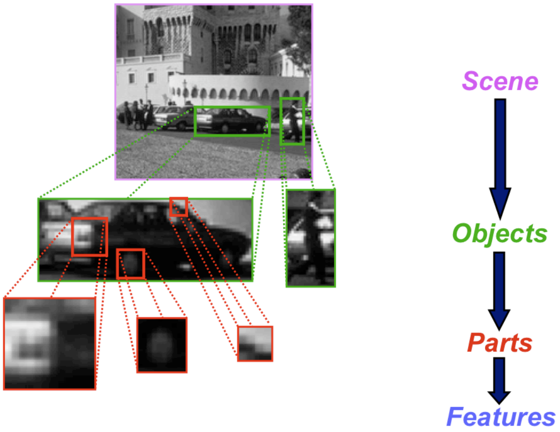


Figure from Sudderth et al. [Sudderth et al. 2007]

Summary

- Dirichlet process is “just” a glorified Dirichlet distribution.
- Draws from a DP are probability measures consisting of a weighted sum of point masses.
- Many representations: Pólya urn scheme, Chinese restaurant process, stick-breaking construction, normalized gamma process.
- DP mixture models are mixture models with countably infinite number of components.
- Important underpinning concepts: de Finetti’s Theorem, Kolmogorov Consistency Theorem.
- I have not delved into inference.

Tutorials on Nonparametric Bayes

- Zoubin Gharamani, UAI 2005.
- Michael Jordan, NIPS 2005.
- Volker Tresp, ICML nonparametric Bayes workshop 2006.
- Workshop on Bayesian Nonparametric Regression, Cambridge, July 2007.
- My Machine Learning Summer School 2007 tutorial and practical course.

Outline

Bayesian Nonparametric Modelling

Dirichlet Processes

Representations of Dirichlet Processes

Some Applications of Dirichlet Processes

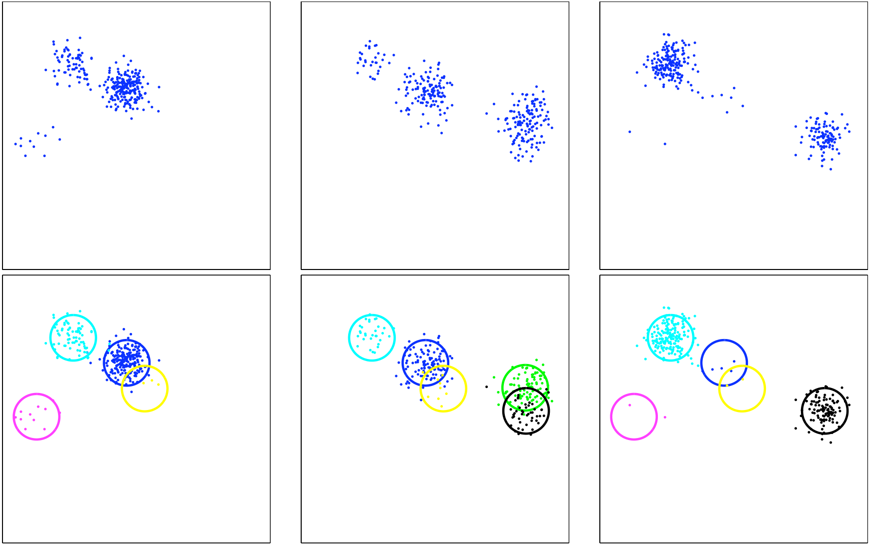
Hierarchical Dirichlet Processes

- Grouped Clustering
- Hierarchical Dirichlet Processes
- Representations

Nested and Dependent Dirichlet Processes

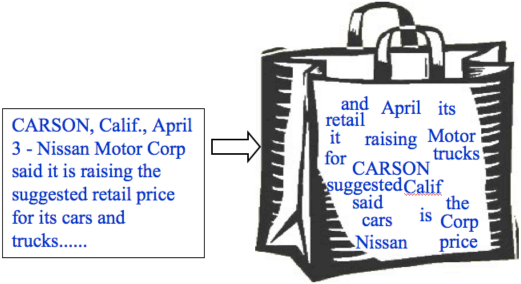
Indian Buffet Processes

Grouped Clustering



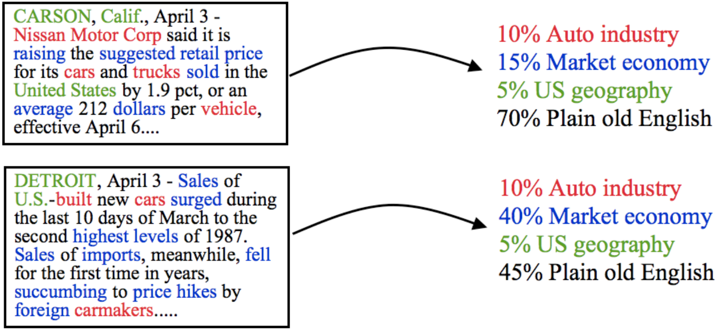
Document Topic Modelling

- Information retrieval: finding useful information from large collections of documents.
- Example: Google, CiteSeer, Amazon...
- Model documents as “bags of words”.



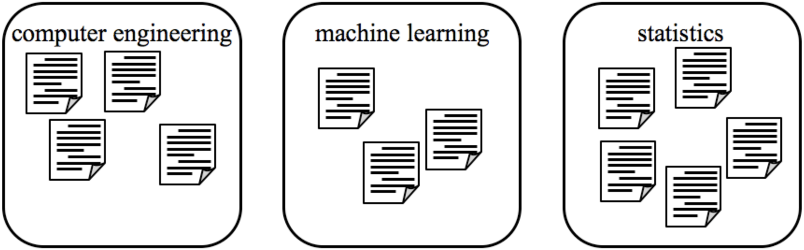
Document Topic Modelling

- We model documents as coming from an underlying set of topics.
  - Summarize documents.
  - Document/query comparisons.
  - Do not know the number of topics a priori—use DP mixtures somehow.
  - But: topics have to be shared across documents...



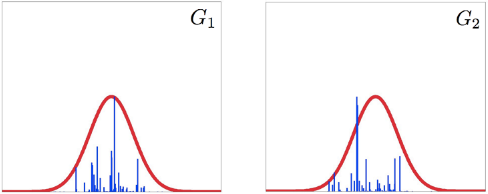
Document Topic Modelling

- Share topics across documents in a collection, and across different collections.
- More sharing within collections than across.
- Use DP mixture models as we do not know the number of topics a priori.



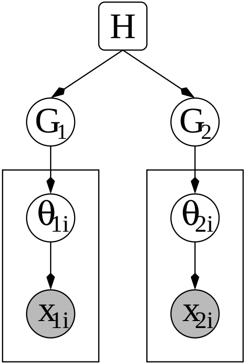
Hierarchical Dirichlet Processes

- Use a DP mixture for each group.



- Unfortunately there is no sharing of clusters across different groups because *H* is smooth.
- Solution: make the base distribution *H* discrete.
- Put a DP prior on the common base distribution.

[Teh et al. 2006]

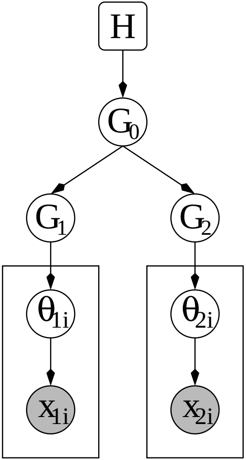


Hierarchical Dirichlet Processes

- A hierarchical Dirichlet process:

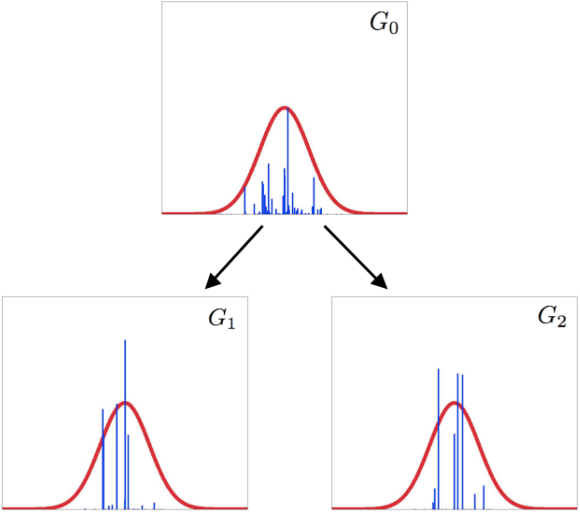
$$\begin{aligned} G_0 &\sim \mathrm{DP}(\alpha_0, H) \\ G_1, G_2 | G_0 &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

- Extension to other hierarchies is straightforward.



Hierarchical Dirichlet Processes

- Making G_0 discrete forces shared cluster between G_1 and G_2.



Stick-breaking Construction

- We shall assume the following HDP hierarchy:

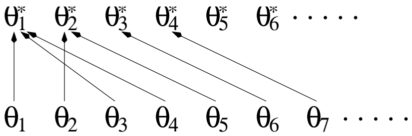
$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_j | G_0 &\sim \mathrm{DP}(\alpha, G_0) \quad \text{for } j = 1, \ldots, J
\end{aligned}
$$

- The stick-breaking construction for the HDP is:

$$
\begin{aligned}
G_0 &= \textstyle\sum_{k=1}^{\infty} \pi_{0k} \delta_{\theta_k^*} & \theta_k^* &\sim H \\
\pi_{0k} &= \beta_{0k} \textstyle\prod_{l=1}^{k-1} (1 - \beta_{0l}) & \beta_{0k} &\sim \mathrm{Beta}\left(1, \gamma\right) \\
G_j &= \textstyle\sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^*} & & \\
\pi_{jk} &= \beta_{jk} \textstyle\prod_{l=1}^{k-1} (1 - \beta_{jl}) & \beta_{jk} &\sim \mathrm{Beta}\left(\alpha\beta_{0k}, \alpha(1 - \textstyle\sum_{l=1}^{k} \beta_{0l})\right)
\end{aligned}
$$

Hierarchical Pòlya Urn Scheme

- Let $G \sim \mathrm{DP}(\alpha, H)$.
- We can visualize the Pòlya urn scheme as follows:



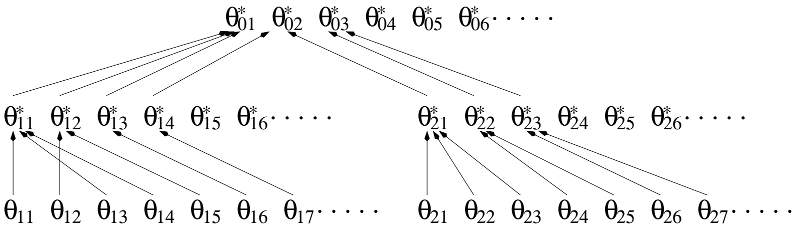
where the arrows denote to which θ_k^* each θ_i was assigned and

$$\theta_1, \theta_2, \ldots \sim G \text{ i.i.d.}$$
$$\theta_1^*, \theta_2^*, \ldots \sim H \text{ i.i.d.}$$

(but $\theta_1, \theta_2, \ldots$ are not independent of $\theta_1^*, \theta_2^*, \ldots$).

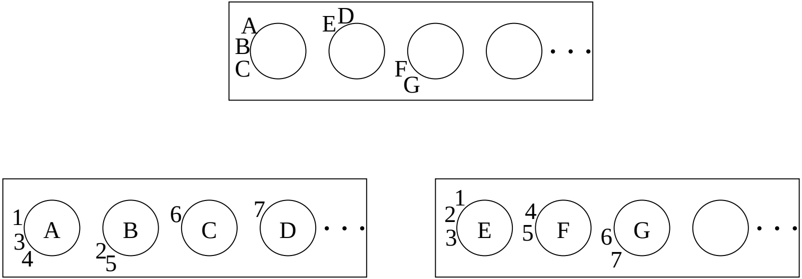
Hierarchical Pòlya Urn Scheme

- Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.
- The hierarchical Pòlya urn scheme to generate draws from G_1, G_2:



Chinese Restaurant Franchise

- Let $G_0 \sim \mathrm{DP}(\gamma, H)$ and $G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0)$.
- The Chinese restaurant franchise describes the clustering of data items in the hierarchy:



Outline

Bayesian Nonparametric Modelling

Dirichlet Processes

Representations of Dirichlet Processes

Some Applications of Dirichlet Processes

Hierarchical Dirichlet Processes

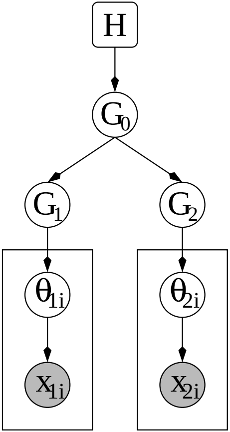
Nested and Dependent Dirichlet Processes

Indian Buffet Processes

Nested Dirichlet Processes

- The HDP assumes that data group structure is observed.
- The group structure may not be known in practice, even if there is prior belief in some group structure.
- Even if known, we may still believe that some groups are more similar to each other than to other groups.
- We can *cluster groups* using a second level of mixture models.
- Using a second DP mixture to model this leads to the *nested Dirichlet process*.

[Rodríguez et al. 2006]



Nested Dirichlet Processes

- Cluster groups. Each group j belongs to cluster k_j:

$$k_j \sim \pi \qquad\qquad \pi \sim \mathrm{GEM}(\alpha)$$

- Group j inherits the DP from cluster k_j:

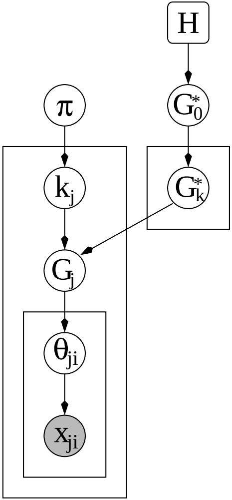
$$G_j = G^*_{k_j}$$

- Place a HDP prior on $\{G^*_k\}$ (not crucial):

$$G^*_k \sim \mathrm{DP}(\beta, G^*_0) \qquad\qquad G^*_0 \sim \mathrm{DP}(\gamma, H)$$

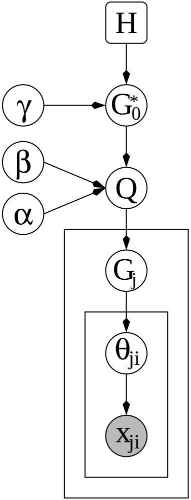
- Data:

$$x_{ji} \sim F(\theta_{ji}) \qquad\qquad \theta_{ji} \sim G_j$$



Nested Dirichlet Processes

$$
\begin{aligned}
G_0^* &\sim \mathrm{DP}(\gamma, H) \\
Q &\sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, G_0^*)) \\
G_j &\sim Q \\
\theta_{ji} &\sim G_j \\
x_{ji} &\sim F(\theta_{ji})
\end{aligned}
$$



Dependent Dirichlet Processes

- The HDP induces a straightforward dependency among groups.
- What if the data is smoothly varying across some spatial or temporal domain?
  - Topic modelling: topic popularity and composition can both change slowly as time passes.
  - Haplotype inference: haplotype occurrence can change smoothly as function of geography.
- a dependent Dirichlet process is a stochastic process $\{G_t\}$ indexed by t (space or time), such that each $G_t \sim \mathrm{DP}(\alpha, H)$ and if t, t' are neighbouring points, G_t and $G_{t'}$ should be "similar" to each other.
- Simple example:

$$\pi \sim \mathrm{GEM}(\alpha) \qquad\qquad (\theta^*_{tk}) \sim \mathrm{GP}(\mu, \Sigma) \quad \text{for each } k$$

$$G_t = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_{tk}}$$

Outline

Bayesian Nonparametric Modelling

Dirichlet Processes

Representations of Dirichlet Processes

Some Applications of Dirichlet Processes

Hierarchical Dirichlet Processes

Nested and Dependent Dirichlet Processes

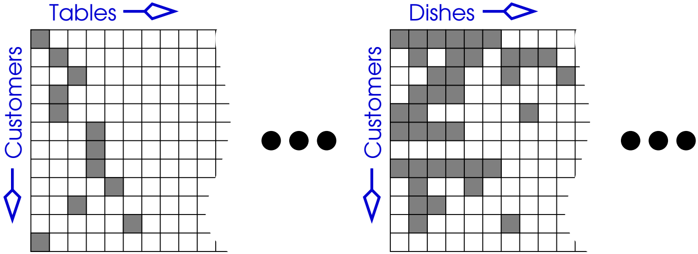
Indian Buffet Processes

Beyond Clustering

- Dirichlet processes are nonparametric models of clustering.
- Can nonparametric models go beyond clustering to describe data in more expressive ways?
  - Hierarchical (e.g. taxonomies)?
  - Distributed (e.g. multiple causes)?

Indian Buffet Processes

- The *Indian Buffet Process* (IBP) is akin to the Chinese restaurant process but describes each customer with a binary vector instead of cluster.
- Generating from an IBP:
  - Parameter α.
  - First customer picks Poisson(α) dishes to eat.
  - Subsequent customer i picks dish k with probability $\frac{n_k}{i}$; and picks Poisson($\frac{\alpha}{i}$) new dishes.



Indian Buffet Processes

- The IBP is infinitely exchangeable, though this is much harder to see.
- De Finetti's Theorem again states that there is some random measure underlying the IBP.
- This random measure is the Beta process.

[Griffiths and Ghahramani 2006, Thibaux and Jordan 2007]

Beta Processes

- A *beta process* $B \sim \mathrm{BP}(c, \alpha H)$ is a random discrete measure with form:

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

where the points $P = \{(\theta_1^*, \mu_1), (\theta_2^*, \mu_2), \ldots\}$ are spikes in a 2D Poisson process with base measure:

$$\alpha c \mu^{-1} (1-\mu)^{c-1} d\mu H(d\theta)$$

- The beta process with $c = 1$ is the de Finetti measure for the IBP. When $c \neq 1$ we have a two parameter generalization of the IBP.
- This is an example of a *completely random measure*.
- A beta process *does not* have Beta distributed marginals.

[Hjort 1990]

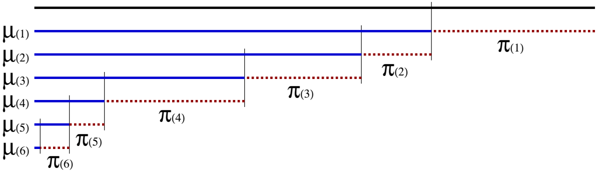
Stick-breaking Construction for Beta Processes

- When $c = 1$ it was shown that the following generates a draw of B:

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \mu_k = (1 - \beta_k) \prod_{i=1}^{k-1} (1 - \beta_i) \qquad \theta_k^* \sim H$$

$$B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

- The above is the complement of the stick-breaking construction for DPs!



[Teh et al. 2007]

Indian Buffet Processes

Applications of Indian Buffet Processes.

- The IBP can be used in concert with different likelihood models in a variety of applications.

$$Z \sim \mathrm{IBP}(\alpha) \qquad\qquad X \sim F(Z, Y)$$

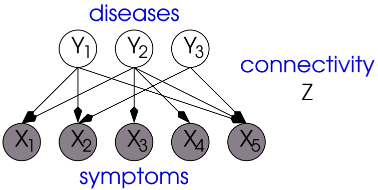
$$Y \sim H \qquad\qquad p(Z, Y|X) = \frac{p(Z, Y)p(X|Z, Y)}{p(X)}$$

- Latent factor models for distributed representation [Griffiths and Ghahramani 2005].
- Matrix factorization for collaborative filtering [Meeds et al 2007].
- Latent causal discovery for medical diagnostics [Wood et al 2006].
- Protein complex discovery [Chu et al 2006].
- Psychological choice behaviour [Görür and Rasmussen 2006].
- Independent Components Analysis
  [Knowles and Ghahramani 2007, Teh et al. 2007].

Indian Buffet Processes

Application: causal discovery [Wood et al 2006].

- Causal model of patient symptoms and diseases.



Noisy-or observations:

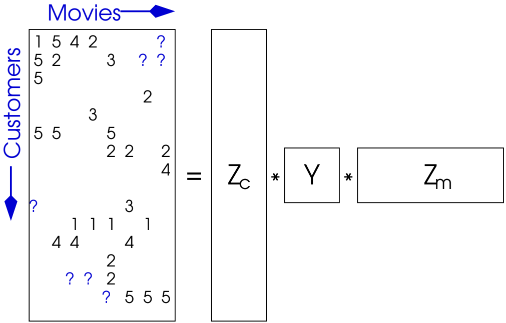
$$p(X_{it} = 1 | Y, Z) = 1 - (1 - \epsilon_i) \prod_k (1 - \lambda_{ik})^{z_{ik} y_{kt}}$$

- Usually given model and the task is to infer diseases Y given X.
- Given sets of patient symptoms X, we can learn the disease causes of these symptoms by learning both Y and Z.

Indian Buffet Processes

Application: collaborative filtering [Meeds et al 2007].

- Model how customers like movies in terms of binary features Z_c, Z_m and interaction matrix Y.



Indian Buffet Processes

Application: Independent Components Analysis

- Each image X_i is a linear combination of sparse features:

$$X_i = \sum_k \Lambda_k y_{ik}$$

where y_{ik} is activity of feature k with sparse prior. One possibility is a mixture of a Gaussian and a point mass at 0:

$$y_{ik} = z_{ik} a_{ik} \qquad a_{ik} \sim \mathcal{N}(0,1) \qquad Z \sim \text{IBP}(\alpha)$$

- An ICA model with infinite number of features.

[Knowles and Ghahramani 2007, Teh et al. 2007]

Next Week

- Infinite hidden Markov models [Beal et al. 2002] (NIPS).
- Dirichlet diffusion trees [Neal 2003] (Valencia).
- Nested Chinese restaurant processes [Blei et al. 2004] (NIPS).
- Infinite relational model [Kemp et al. 2006] (AAAI)?

Bibliography I

Dirichlet Processes and Beyond in Machine Learning

Dirichlet Processes were first introduced by [Ferguson 1973], while [Antoniak 1974] further developed DPs as well as introduce the mixture of DPs. [Blackwell and MacQueen 1973] showed that the Pólya urn scheme is exchangeable with the DP being its de Finetti measure. Further information on the Chinese restaurant process can be obtained at [Aldous 1985, Pitman 2002]. The DP is also related to Ewens' Sampling Formula [Ewens 1972]. [Sethuraman 1994] gave a constructive definition of the DP via a stick-breaking construction. DPs were rediscovered in the machine learning community by [Neal 1992, Rasmussen 2000].

Hierarchical Dirichlet Processes (HDPs) were first developed by [Teh et al. 2006], although an aspect of the model was first discussed in the context of infinite hidden Markov models [Beal et al. 2002]. HDPs and generalizations have been applied across a wide variety of fields.

Dependent Dirichlet Processes are sets of coupled distributions over probability measures, each of which is marginally DP [MacEachern et al. 2001]. A variety of dependent DPs have been proposed in the literature since then [Srebro and Roweis 2005, Griffin 2007, Caron et al. 2007]. The infinite mixture of Gaussian processes of [Rasmussen and Ghahramani 2002] can also be interpreted as a dependent DP.

Indian Buffet Processes (IBPs) were first proposed in [Griffiths and Ghahramani 2006], and extended to a two-parameter family in [Griffiths et al. 2007b]. [Thibaux and Jordan 2007] showed that the de Finetti measure for the IBP is the beta process of [Hjort 1990], while [Teh et al. 2007] gave a stick-breaking construction and developed efficient slice sampling inference algorithms for the IBP.

Nonparametric Tree Models are models that use distributions over trees that are consistent and exchangeable. [Blei et al. 2004] used a nested CRP to define distributions over trees with a finite number of levels. [Neal 2001, Neal 2003] defined Dirichlet diffusion trees, which are binary trees produced by a fragmentation process. [Teh et al. 2008] used Kingman's coalescent [Kingman 1982b, Kingman 1982a] to produce random binary trees using a coalescent process. [Roy et al. 2007] proposed annotated hierarchies, using tree-consistent partitions first defined in [Heller and Ghahramani 2005] to model both relational and featural data.

Markov chain Monte Carlo Inference algorithms are the dominant approaches to inference in DP mixtures. [Neal 2000] is a good review of algorithms based on Gibbs sampling in the CRP representation. Algorithm 8 in [Neal 2000] is still one of the best algorithms based on simple local moves. [Ishwaran and James 2001] proposed blocked Gibbs sampling in the stick-breaking representation instead due to the simplicity in implementation. This has been further explored in [Porteous et al. 2006]. Since then there has been proposals for better MCMC samplers based on proposing larger moves in a Metropolis-Hastings framework [Jain and Neal 2004, Liang et al. 2007a], as well as sequential Monte Carlo [Fearnhead 2004, Mansingkha et al. 2007].

Other Approximate Inference Methods have also been proposed for DP mixture models. [Blei and Jordan 2006] is the first variational Bayesian approximation, and is based on a truncated stick-breaking representation. [Kurihara et al. 2007] proposed an improved VB approximation based on a better truncation technique, and using KD-trees for extremely efficient inference in large scale applications. [Kurihara et al. 2007] studied improved VB approximations based on integrating out the stick-breaking weights. [Minka and Ghahramani 2003] derived an expectation propagation based algorithm. [Heller and Ghahramani 2005] derived tree-based approximation which can be seen as a Bayesian hierarchical clustering algorithm. [Daume III 2007] developed admissible search heuristics to find MAP clusterings in a DP mixture model.

Computer Vision and Image Processing. HDPs have been used in object tracking [Fox et al. 2006, Fox et al. 2007b, Fox et al. 2007a]. An extension called the transformed Dirichlet process has been used in scene analysis [Sudderth et al. 2006b, Sudderth et al. 2006a, Sudderth et al. 2007], a related extension has been used in fMRI image analysis [Kim and Smyth 2007, Kim 2007]. An extension of the infinite hidden Markov model called the nonparametric hidden Markov tree has been introduced and applied to image denoising [Kivinen et al. 2007].
Natural Language Processing. HDPs are essential ingredients in defining nonparametric context free grammars [Liang et al. 2007b, Finkel et al. 2007]. [Johnson et al. 2007] defined adaptor grammars, which is a framework generalizing both probabilistic context free grammars as well as a variety of nonparametric models including DPs and HDPs. DPs and HDPs have been used in information retrieval [Cowans 2004], word segmentation [Goldwater et al. 2006b], word morphology modelling [Goldwater et al. 2006a], coreference resolution [Haghighi and Klein 2007], topic modelling [Blei et al. 2004, Teh et al. 2006, Li et al. 2007]. An extension of the HDP called the hierarchical Pitman-Yor process has been applied to language modelling [Teh 2006a, Teh 2006b, Goldwater et al. 2006a].[Savova et al. 2007] used annotated hierarchies to construct syntactic hierarchies. Theses on nonparametric methods in NLP include [Cowans 2006, Goldwater 2006].
Other Applications. Applications of DPs, HDPs and infinite HMMs in bioinformatics include [Xing et al. 2004, Xing et al. 2006, Xing et al. 2007, Xing and Sohn 2007a, Xing and Sohn 2007b]. DPs have been applied in relational learning [Shafto et al. 2006, Kemp et al. 2006, Xu et al. 2006], spike sorting [Wood et al. 2006a, Görür 2007]. The HDP has been used in a cognitive model of categorization [Griffiths et al. 2007a]. IBPs have been applied to infer hidden causes [Wood et al. 2006b], in a choice model [Görür et al. 2006], to modelling dyadic data [Meeds et al. 2007], to overlapping clustering [Heller and Ghahramani 2007], and to matrix factorization [Wood and Griffiths 2006].

References I

Aldous, D. (1985).
Exchangeability and related topics.
In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin.

Antoniak, C. E. (1974).
Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.
Annals of Statistics, 2(6):1152–1174.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002).
The infinite hidden Markov model.
In *Advances in Neural Information Processing Systems*, volume 14.

Blackwell, D. and MacQueen, J. B. (1973).
Ferguson distributions via Pólya urn schemes.
Annals of Statistics, 1:353–355.

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004).
Hierarchical topic models and the nested Chinese restaurant process.
In *Advances in Neural Information Processing Systems*, volume 16.

Blei, D. M. and Jordan, M. I. (2006).
Variational inference for Dirichlet process mixtures.
Bayesian Analysis, 1(1):121–144.

Caron, F., Davy, M., and Doucet, A. (2007).
Generalized Polya urn for time-varying Dirichlet process mixtures.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 23.

References II

Cowans, P. (2004).
Information retrieval using hierarchical Dirichlet processes.
In *Proceedings of the Annual International Conference on Research and Development in Information Retrieval*, volume 27, pages 564–565.

Cowans, P. (2006).
Probabilistic Document Modelling.
PhD thesis, University of Cambridge.

Daume III, H. (2007).
Fast search for Dirichlet process mixture models.
In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Ewens, W. J. (1972).
The sampling theory of selectively neutral alleles.
Theoretical Population Biology, 3:87–112.

Fearnhead, P. (2004).
Particle filters for mixture models with an unknown number of components.
Statistics and Computing, 14:11–21.

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
Annals of Statistics, 1(2):209–230.

Finkel, J. R., Grenager, T., and Manning, C. D. (2007).
The infinite tree.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

References III

Fox, E. B., Choi, D. S., and Willsky, A. S. (2006).
Nonparametric Bayesian methods for large scale multi-target tracking.
In *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, volume 40.

Fox, E. B., Sudderth, E. B., Choi, D. S., and Willsky, A. S. (2007a).
Tracking a non-cooperative maneuvering target using hierarchical Dirichlet processes.
In *Proceedings of the Adaptive Sensor Array Processing Conference*.

Fox, E. B., Sudderth, E. B., and Willsky, A. S. (2007b).
Hierarchical Dirichlet processes for tracking maneuvering targets.
In *Proceedings of the International Conference on Information Fusion*.

Goldwater, S. (2006).
Nonparametric Bayesian Models of Lexical Acquisition.
PhD thesis, Brown University.

Goldwater, S., Griffiths, T., and Johnson, M. (2006a).
Interpolating between types and tokens by estimating power-law generators.
In *Advances in Neural Information Processing Systems*, volume 18.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006b).
Contextual dependencies in unsupervised word segmentation.
In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Görür, D. (2007).
Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning.
PhD thesis, Technischen Universität Berlin.

References IV

Görür, D., Jäkel, F., and Rasmussen, C. E. (2006).
A choice model with infinitely many latent features.
In *Proceedings of the International Conference on Machine Learning*, volume 23.

Griffin, J. E. (2007).
The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference.
Technical report, Department of Statistics, University of Warwick.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a).
Unifying rational models of categorization via the hierarchical Dirichlet process.
In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 29.

Griffiths, T. L. and Ghahramani, Z. (2006).
Infinite latent feature models and the Indian buffet process.
In *Advances in Neural Information Processing Systems*, volume 18.

Griffiths, T. L., Ghahramani, Z., and Sollich, P. (2007b).
Bayesian nonparametric latent feature models (with discussion and rejoinder).
In *Bayesian Statistics*, volume 8.

Haghighi, A. and Klein, D. (2007).
Unsupervised coreference resolution in a nonparametric Bayesian model.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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

References V

Heller, K. A. and Ghahramani, Z. (2007).
A nonparametric Bayesian approach to modeling overlapping clusters.
In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Hjort, N. L. (1990).
Nonparametric Bayes estimators based on beta processes in models for life history data.
Annals of Statistics, 18(3):1259–1294.

Ishwaran, H. and James, L. F. (2001).
Gibbs sampling methods for stick-breaking priors.
Journal of the American Statistical Association, 96(453):161–173.

Jain, S. and Neal, R. M. (2004).
A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.
Technical report, Department of Statistics, University of Toronto.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007).
Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models.
In *Advances in Neural Information Processing Systems*, volume 19.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006).
Learning systems of concepts with an infinite relational model.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 21.

Kim, S. (2007).
Learning Hierarchical Probabilistic Models with Random Effects with Applications to Time-series and Image Data.
PhD thesis, Information and Computer Science, University of California at Irvine.

References VI

Kim, S. and Smyth, P. (2007).
Hierarchical dirichlet processes with random effects.
In *Advances in Neural Information Processing Systems*, volume 19.

Kingman, J. F. C. (1982a).
The coalescent.
Stochastic Processes and their Applications, 13:235–248.

Kingman, J. F. C. (1982b).
On the genealogy of large populations.
Journal of Applied Probability, 19:27–43.
Essays in Statistical Science.

Kivinen, J., Sudderth, E., and Jordan, M. I. (2007).
Image denoising with nonparametric hidden Markov trees.
In *International Conference on Image Processing*.

Knowles, D. and Ghahramani, Z. (2007).
Infinite sparse factor analysis and infinite independent components analysis.
In *7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, Lecture Notes in Computer Science Series (LNCS). Springer.

Kurihara, K., Welling, M., and Vlassis, N. (2007).
Accelerated variational DP mixture models.
In *Advances in Neural Information Processing Systems*, volume 19.

Li, W., Blei, D., and McCallum, A. (2007).
Nonparametric Bayes pachinko allocation.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

References VII

Liang, P., Jordan, M. I., and Taskar, B. (2007a).
A permutation-augmented sampler for Dirichlet process mixture models.
In *Proceedings of the International Conference on Machine Learning*.

Liang, P., Petrov, S., Jordan, M. I., and Klein, D. (2007b).
The infinite PCFG using hierarchical Dirichlet processes.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

MacEachern, S., Kottas, A., and Gelfand, A. (2001).
Spatial nonparametric Bayesian models.
Technical Report 01-10, Institute of Statistics and Decision Sciences, Duke University.
http://ftp.isds.duke.edu/WorkingPapers/01-10.html.

Mansingkha, V. K., Roy, D. M., Rifkin, R., and Tenenbaum, J. B. (2007).
AClass: An online algorithm for generative classification.
In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007).
Modeling dyadic data with binary latent factors.
In *Advances in Neural Information Processing Systems*, volume 19.

Minka, T. P. and Ghahramani, Z. (2003).
Expectation propagation for infinite mixtures.
Presented at NIPS2003 Workshop on Nonparametric Bayesian Methods and Infinite Models.

Neal, R. M. (1992).
Bayesian mixture modeling.
In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, volume 11, pages 197–211.

References VIII

Neal, R. M. (2000).
Markov chain sampling methods for Dirichlet process mixture models.
Journal of Computational and Graphical Statistics, 9:249–265.

Neal, R. M. (2001).
Defining priors for distributions using Dirichlet diffusion trees.
Technical Report 0104, Department of Statistics, University of Toronto.

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

Perman, M., Pitman, J., and Yor, M. (1992).
Size-biased sampling of Poisson point processes and excursions.
Probability Theory and Related Fields, 92(1):21–39.

Pitman, J. (2002).
Combinatorial stochastic processes.
Technical Report 621, Department of Statistics, University of California at Berkeley.
Lecture notes for St. Flour Summer School.

Pitman, J. and Yor, M. (1997).
The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.
Annals of Probability, 25:855–900.

Porteous, I., Ihler, A., Smyth, P., and Welling, M. (2006).
Gibbs sampling for (Coupled) infinite mixture models in the stick-breaking representation.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 22.

References IX

Rasmussen, C. E. (2000).
The infinite Gaussian mixture model.
In *Advances in Neural Information Processing Systems*, volume 12.

Rasmussen, C. E. and Ghahramani, Z. (2001).
Occam's razor.
In *Advances in Neural Information Processing Systems*, volume 13.

Rasmussen, C. E. and Ghahramani, Z. (2002).
Infinite mixtures of Gaussian process experts.
In *Advances in Neural Information Processing Systems*, volume 14.

Rasmussen, C. E. and Williams, C. K. I. (2006).
Gaussian Processes for Machine Learning.
MIT Press.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2006).
The nested Dirichlet process.
Technical Report 2006-19, Institute of Statistics and Decision Sciences, Duke University.

Roy, D. M., Kemp, C., Mansinghka, V., and Tenenbaum, J. B. (2007).
Learning annotated hierarchies from relational data.
In *Advances in Neural Information Processing Systems*, volume 19.

Savova, V., Roy, D., Schmidt, L., and Tenenbaum, J. B. (2007).
Discovering syntactic hierarchies.
In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 29.

References X

Sethuraman, J. (1994).
A constructive definition of Dirichlet priors.
Statistica Sinica, 4:639–650.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., and Tenenbaum, J. B. (2006).
Learning cross-cutting systems of categories.
In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 28.

Srebro, N. and Roweis, S. (2005).
Time-varying topic models using dependent Dirichlet processes.
Technical Report UTML-TR-2005-003, Department of Computer Science, University of Toronto.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2006a).
Depth from familiar objects: A hierarchical model for 3D scenes.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2006b).
Describing visual scenes using transformed Dirichlet processes.
In *Advances in Neural Information Processing Systems*, volume 18.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2007).
Describing visual scenes using transformed objects and parts.
To appear in the International Journal of Computer Vision.

Teh, Y. W. (2006a).
A Bayesian interpretation of interpolated Kneser-Ney.
Technical Report TRA2/06, School of Computing, National University of Singapore.

References XI

Teh, Y. W. (2006b).
A hierarchical Bayesian language model based on Pitman-Yor processes.
In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

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

Teh, Y. W., Görür, D., and Ghahramani, Z. (2007).
Stick-breaking construction for the Indian buffet process.
In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical Dirichlet processes.
Journal of the American Statistical Association, 101(476):1566–1581.

Thibaux, R. and Jordan, M. I. (2007).
Hierarchical beta processes and the Indian buffet process.
In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Wood, F., Goldwater, S., and Black, M. J. (2006a).
A non-parametric Bayesian approach to spike sorting.
In *Proceedings of the IEEE Conference on Engineering in Medicine and Biologicial Systems*, volume 28.

Wood, F. and Griffiths, T. L. (2006).
Particle filtering for nonparametric Bayesian matrix factorization.
In *Advances in Neural Information Processing Systems*, volume 18.

References XII

Wood, F., Griffiths, T. L., and Ghahramani, Z. (2006b).
A non-parametric Bayesian method for inferring hidden causes.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 22.

Xing, E., Sharan, R., and Jordan, M. (2004).
Bayesian haplotype inference via the dirichlet process.
In *Proceedings of the International Conference on Machine Learning*, volume 21.

Xing, E. P., Jordan, M. I., and Roded, R. (2007).
Bayesian haplotype inference via the Dirichlet process.
Journal of Computational Biology, 14(3):267–284.

Xing, E. P. and Sohn, K. (2007a).
Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space.
Bayesian Analysis, 2(2).

Xing, E. P. and Sohn, K. (2007b).
A nonparametric Bayesian approach for haplotype reconstruction from single and multi-population data.
Technical Report CMU-MLD 07-107, Carnegie Mellow University.

Xing, E. P., Sohn, K., Jordan, M. I., and Teh, Y. W. (2006).
Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture.
In *Proceedings of the International Conference on Machine Learning*, volume 23.

Xu, Z., Tresp, V., Yu, K., and Kriegel, H.-P. (2006).
Infinite hidden relational models.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 22.