An Introduction to Bayesian Nonparametric Modelling

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

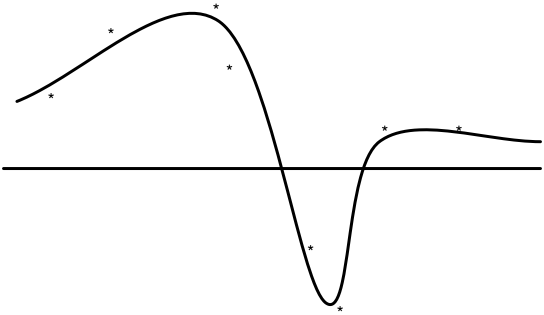
September, 2009 / MLSS Cambridge

Outline

Outline

Regression with Basis Functions

- Supervised learning of a function $f^* : \mathbb{X} \rightarrow \mathbb{Y}$ from training data $\{x_i, y_i\}_{i=1}^{n}$.



Regression with Basis Functions

- Assume a set of basis functions $\phi_1, \ldots, \phi_K$ and parametrize a function:

$$f(x; \mathbf{w}) = \sum_{k=1}^{K} w_k \phi_k(x)$$

  Parameters $\mathbf{w} = \{w_1, \ldots, w_K\}$.

- Find optimal parameters

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \left| y_i - f(x_i; \mathbf{w}) \right|^2 = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \left| y_i - \textstyle\sum_{k=1}^{K} w_k \phi_k(x_i) \right|^2$$

- We will be Bayesian in this lecture, so we need to rephrase using probabilistic model with priors on parameters:

$$y_i | x_i, \mathbf{w} = f(x_i; \mathbf{w}) + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$w_k \sim \mathcal{N}(0, \tau^2)$$

- Computer posterior $p(\mathbf{w} | \{x_i, y_i\})$.

Regression with Basis Functions

$$f(x; \mathbf{w}) = \sum_{k=1}^{K} w_k \phi_k(x)$$

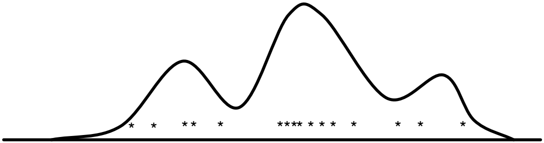
- What basis functions to use?
- How many basis functions to use?
- Do we really believe that the true $f^*(x)$ can be expressed as $f^*(x) = f(x; \mathbf{w}^*)$ for some $\mathbf{w}^*$?

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Do we believe the noise process is Gaussian?

Density Estimation with Mixture Models

- Unsupervised learning of a density $f^*(x)$ from training samples $\{x_i\}$.



- Perhaps use an exponential family distribution, e.g. Gaussian?

$$\mathcal{N}(x;\mu,\Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

  Unimodal, restrictive shape, light tail...

- Use a mixture model instead,

$$f(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x;\mu_k,\Sigma_k)$$

- Do we believe that the true density is a mixture of K components?
- How many mixture components to use?

Latent Variable Modelling

- Say we have n vector observations $x_1, \ldots, x_n$.
- Model each observation as a linear combination of K latent sources:

$$x_i = \sum_{k=1}^{K} \Lambda_k y_{ik} + \epsilon_i$$

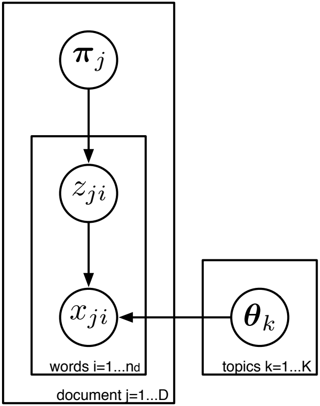
  y_{ik}: activity of source k in datum i.
  Λ_k: basis vector describing effect of source k.
- Examples include principle components analysis, factor analysis, independent components analysis.
- How many sources are there?
- Do we believe that K sources is sufficient to explain all our data?
- What prior distribution should we use for sources?

Topic Modelling with Latent Dirichlet Allocation

- Infer topics from a document corpus, topics being sets of words that tend to co-occur together.
- Using (Bayesian) latent Dirichlet allocation:

$$\begin{aligned}
\boldsymbol{\pi}_j &\sim \text{Dirichlet}(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\tfrac{\beta}{W}, \ldots, \tfrac{\beta}{W}) \\
z_{ji} | \boldsymbol{\pi}_j &\sim \text{Multinomial}(\boldsymbol{\pi}_j) \\
x_{ji} | z_{ji}, \boldsymbol{\theta}_{z_{ji}} &\sim \text{Multinomial}(\boldsymbol{\theta}_{z_{ji}})
\end{aligned}$$

- How many topics can we find from the corpus?



Outline

Modelling Data

- Models are almost never correct for real world data.
- How do we deal with model misfit?
  - Quantify closeness to true model, and optimality of fitted model;
  - Model selection or averaging;
  - Increase the flexibility of your model class.
- Bayesian nonparametrics are good solutions from the second and third perspectives.

Model Selection and Model Averaging

- Data $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.
- Model M_k parametrized by θ_k, for $k = 1, 2, \ldots$.
- 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 and averaging:

$$M = \operatorname*{argmax}_{M_k} p(\mathbf{x}|M_k) \quad \text{or} \quad p(k, \theta_k|\mathbf{x}) = \frac{p(k)p(\theta_k|M_k)p(\mathbf{x}|\theta_k, M_k)}{\sum_{k'} p(k')p(\theta_{k'}|M_{k'})p(\mathbf{x}|\theta_{k'}, M_{k'})}$$

- Model selection and averaging is to prevent overfitting and underfitting, and are usually expense to compute.
- But reasonable and proper Bayesian methods should not overfit anyway [Rasmussen and Ghahramani 2001].

Nonparametric Modelling

- What is a nonparametric model?
  - A parametric model where the number of parameters increases with data;
  - A really large parametric model;
  - A model over infinite dimensional function or measure spaces.
  - A family of distributions that is dense in some large space.
- Why nonparametric models in Bayesian theory of learning?
  - broad class of priors that allows data to “speak for itself”;
  - side-step model selection and averaging.
- How do we deal with the very large parameter spaces?
  - Marginalize out all but a finite number of parameters;
  - 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 $x_1, \ldots, x_n$,

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

  where the parameters are the mean function $m(x)$ and covariance kernel $c(x, y)$.
- Note: a random function f is a stochastic process. It is a collection of random variables $\{f(x)\}_{x \in \mathbb{X}}$ one for each possible input value x.
- Can also be expressed as

$$f(x) = \sum_{k=1}^{K} w_k \phi_k(x) \quad \text{as } K \to \infty.$$

[Rasmussen and Williams 2006]

Posterior and Predictive Distributions

- How do we compute the posterior and predictive distributions?
- Training set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and test input x_{n+1}.
- Out of the (uncountably infinitely) many random variables $\{f(x)\}_{x \in \mathbb{X}}$ making up the GP only $n + 1$ has to do with the data:

$$f(x_1), f(x_2), \ldots, f(x_{n+1})$$

- Training data gives observations $f(x_1) = y_1, \ldots, f(x_n) = y_n$. The predictive distribution of $f(x_{n+1})$ is simply

$$p(f(x_{n+1}) | f(x_1) = y_1, \ldots, f(x_n) = y_n)$$

  which is easy to compute since $f(x_1), \ldots, f(x_{n+1})$ is Gaussian.
- This can be generalized to noisy observations $y_i = f(x_i) + \epsilon_i$ or non-linear effects $y_i \sim D(f(x_i))$ where $D(\theta)$ is a distribution parametrized by θ.

Consistency and Existence

- The definition of Gaussian processes only give finite dimensional marginal distributions of the stochastic process.
- Fortunately these marginal distributions are *consistent*.
  - For every finite set $\mathbf{x} \subset \mathbb{X}$ we have a distinct distribution $p_{\mathbf{x}}([f(x)]_{x \in \mathbf{x}})$. These distributions are said to be consistent if

$$p_{\mathbf{x}}([f(x)]_{x \in \mathbf{x}}) = \int p_{\mathbf{x} \cup \mathbf{y}}([f(x)]_{x \in \mathbf{x} \cup \mathbf{y}}) d[f(x)]_{x \in \mathbf{y}}$$

    for disjoint and finite $\mathbf{x}, \mathbf{y} \subset \mathbb{X}$.
  - The marginal distributions for the GP are consistent because *Gaussians are closed under marginalization*.
- The *Kolmogorov Consistency Theorem* guarantees existence of GPs, i.e. the whole stochastic process $\{f(x)\}_{x \in \mathbb{X}}$.
  - Further information in Peter Orbanz' lectures.

Outline

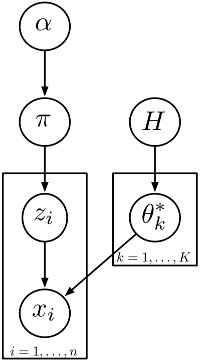
Bayesian Mixture Models

- Let's be Bayesian about mixture models, and place priors over our parameters (and to compute posteriors).
- First, introduce variable z_i indicator which component x_i belongs to.

$$z_i | \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$x_i | z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\mu_k, \Sigma_k)$$

- Second, introduce conjugate priors for parameters:

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K})$$
$$\mu_k, \Sigma_k \sim H = \mathcal{N}\text{-}\mathcal{IW}(0, s, d, \Phi)$$

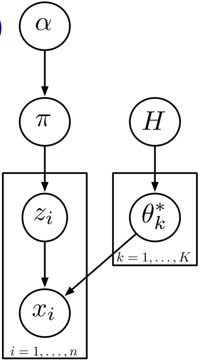


[Rasmussen 2000]

Gibbs Sampling for Bayesian Mixture Models

- All conditional distributions are simple to compute:

$$p(z_i = k|\text{others}) \propto \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)$$
$$\boldsymbol{\pi}|\mathbf{z} \sim \text{Dirichlet}(\tfrac{\alpha}{K} + n_1(\mathbf{z}), \ldots, \tfrac{\alpha}{K} + n_K(\mathbf{z}))$$
$$\mu_k, \Sigma_k|\text{others} \sim \mathcal{N}\text{-}\mathcal{IW}(\nu', s', d', \Phi')$$



- Not as efficient as collapsed Gibbs sampling which integrates out $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$:

$$p(z_i = k|\text{others}) \propto \frac{\frac{\alpha}{K} + n_k(\mathbf{z}_{-i})}{\alpha + n - 1} \times p(x_i|\{x_{i'} : i' \neq i, z_{i'} = k\})$$

- Demo: fm_demointeractive.

Infinite Bayesian Mixture Models

- We will take $K \to \infty$.
- Imagine a very large value of K.
- There are at most $n < K$ occupied components, so most components are *empty*. We can lump these empty components together:

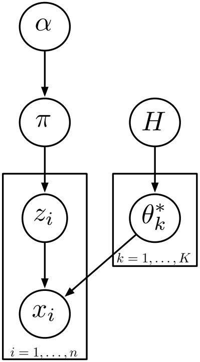
  Occupied clusters:

  $$p(z_i = k|\text{others}) \propto \frac{\frac{\alpha}{K} + n_k(\mathbf{z}_{-i})}{n - 1 + \alpha} p(x_i|\mathbf{x}_k^{-i})$$

  Empty clusters:

  $$p(z_i = k_{\text{empty}}|\mathbf{z}^{-i}) \propto \frac{\alpha \frac{K-K^*}{K}}{n - 1 + \alpha} p(x_i|\{\})$$

- Demo: dpm_demointeractive.



Infinite Bayesian Mixture Models

- We will take $K \to \infty$.
- Imagine a very large value of K.
- There are at most $n < K$ occupied components, so most components are *empty*. We can lump these empty components together:

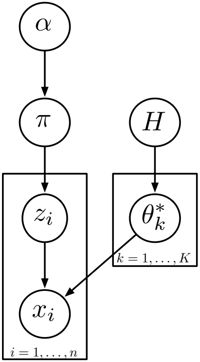
  Occupied clusters:

  $$p(z_i = k|\text{others}) \propto \frac{n_k(\mathbf{z}_{-i})}{n - 1 + \alpha} p(x_i|\mathbf{x}_k^{-i})$$

  Empty clusters:

  $$p(z_i = k_{\text{empty}}|\mathbf{z}^{-i}) \propto \frac{\alpha}{n - 1 + \alpha} p(x_i|\{\})$$

- Demo: dpm_demointeractive.



Infinite Bayesian Mixture Models

- The actual infinite limit of finite mixture models does not make sense: any particular component will get a mixing proportion of 0.
- In the Gibbs sampler we bypassed this by lumping empty clusters together.
- Other better ways of making this infinite limit precise:
  - Look at the prior clustering structure induced by the Dirichlet prior over mixing proportions—*Chinese restaurant process*.
  - Re-order components so that those with larger mixing proportions tend to occur first, before taking the infinite limit—*stick-breaking construction*.
- Both are different views of the *Dirichlet process* (DP).
- DPs can be thought of as infinite dimensional Dirichlet distributions.
- The $K \to \infty$ Gibbs sampler is for DP mixture models.

Outline

A Tiny Bit of Measure Theoretic Probability 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 $\mu : \Sigma \rightarrow [0, \infty]$ 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 $\mu(\Theta) = 1$.
- Given two measure spaces (Θ, Σ) and (Δ, Φ), a function $f : \Theta \rightarrow \Delta$ is *measurable* if $f^{-1}(A) \in \Sigma$ for every $A \in \Phi$.

A Tiny Bit of Measure Theoretic Probability Theory

- If p is a probability measure on (Θ, Σ), a *random variable* X taking values in Δ is simply a measurable function $X : \Theta \to \Delta$.
  - Think of the probability space (Θ, Σ, p) as a black-box random number generator, and X as a function taking random samples in Θ and producing random samples in Δ.
  - The probability of an event $A \in \Phi$ is $p(X \in A) = p(X^{-1}(A))$.
- A *stochastic process* is simply a collection of random variables $\{X_i\}_{i \in \mathbb{I}}$ over the same measure space (Θ, Σ), where $\mathbb{I}$ is an index set.
  - What distinguishes a stochastic process from, say, a graphical model is that $\mathbb{I}$ can be infinite, even uncountably so.
  - This raises issues of how do you even define them and how do you ensure that they can even existence (mathematically speaking).
- Stochastic processes form the core of many Bayesian nonparametric models.
  - Gaussian processes, Poisson processes, gamma processes, Dirichlet processes, beta processes...

Dirichlet Distributions

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

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

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

$$(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\lambda_1, \dots, \lambda_K)$$

  with parameters $(\lambda_1, \dots, \lambda_K)$, if

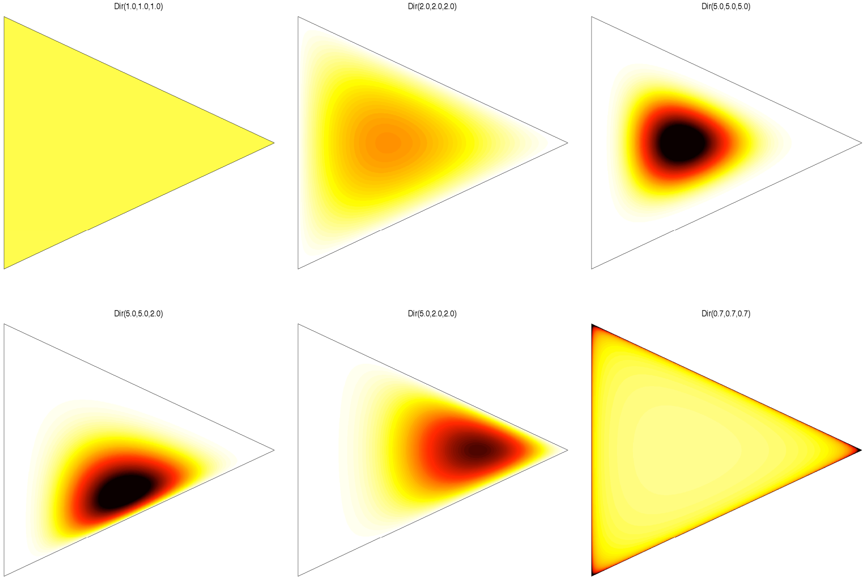
$$p(\pi_1, \dots, \pi_K) = \frac{\Gamma(\sum_k \lambda_k)}{\prod_k \Gamma(\lambda_k)} \prod_{k=1}^{n} \pi_k^{\lambda_k - 1}$$

- Equivalent to normalizing a set of independent gamma variables:

$$(\pi_1, \dots, \pi_K) = \tfrac{1}{\sum_k \gamma_k}(\gamma_1, \dots, \gamma_K)$$

$$\gamma_k \sim \text{Gamma}(\lambda_k) \qquad \text{for } k = 1, \dots, K$$

Dirichlet Distributions

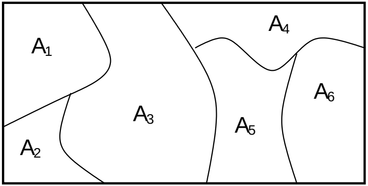


Dirichlet Processes

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

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

  where λ is a base measure.



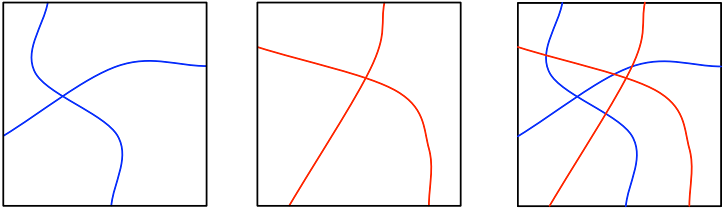
- The above family of distributions is consistent (next slide), and *Kolmogorov Consistency Theorem* can be applied to show existence (but there are technical conditions restricting the generality of the definition).

[Ferguson 1973, Blackwell and MacQueen 1973]

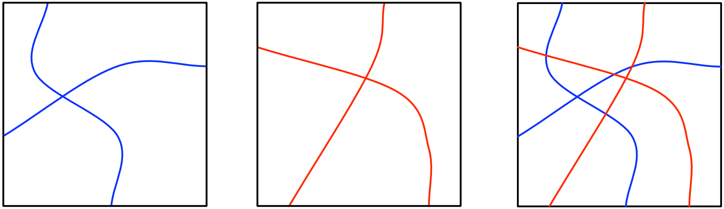
Consistency of Dirichlet Marginals

- If we have two partitions $(A_1, \ldots, A_K)$ and $(B_1, \ldots, B_J)$ of Θ, how do we see if the two Dirichlets are consistent?
- Because Dirichlet variables are normalized gamma variables and sums of gammas are gammas, if $(I_1, \ldots, I_j)$ is a partition of $(1, \ldots, K)$,

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



Consistency of Dirichlet Marginals



- Form the common refinement $(C_1, \ldots, C_L)$ where each C_ℓ is the intersection of some A_k with some B_j. Then:

$$
\begin{aligned}
\text{By definition, } (G(C_1), \ldots, G(C_L)) &\sim \text{Dirichlet}(\lambda(C_1), \ldots, \lambda(C_L)) \\
(G(A_1), \ldots, G(A_K)) &= \left(\textstyle\sum_{C_\ell \subset A_1} G(C_\ell), \ldots, \sum_{C_\ell \subset A_K} G(C_\ell)\right) \\
&\sim \text{Dirichlet}(\lambda(A_1), \ldots, \lambda(A_K)) \\
\text{Similarly, } (G(B_1), \ldots, G(B_J)) &\sim \text{Dirichlet}(\lambda(B_1), \ldots, \lambda(B_J))
\end{aligned}
$$

  so the distributions of $(G(A_1), \ldots, G(A_K))$ and $(G(B_1), \ldots, G(B_J))$ are consistent.

- Demonstration: DPgenerate.

Parameters of Dirichlet Processes

- Usually we split the λ base measure into two parameters $\lambda = \alpha H$:
  - *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 $(A_1, \dots, A_K)$ of Θ:

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

- The first and second moments of the DP:

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

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

where A is any measurable subset of Θ.

Representations of Dirichlet Processes

- Draws from Dirichlet processes will always place all their mass on a countable set of points:

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

  where $\sum_k \pi_k = 1$ and $\theta_k^* \in \Theta$.

  - What is the joint distribution over $\pi_1, \pi_2, \ldots$ and $\theta_1^*, \theta_2^*, \ldots$?

- Since G is a (random) probability measure over Θ, we can treat it as a distribution and draw samples from it. Let

$$\theta_1, \theta_2, \ldots \sim G$$

  be random variables with distribution G.

  - What is the marginal distribution of $\theta_1, \theta_2, \ldots$ with G integrated out?
  - 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. How do these clusters look like?
  - For practical modelling purposes this is sufficient. But is this sufficient to tell us all about G?

Stick-breaking Construction

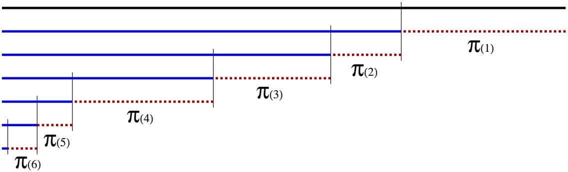
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

- There is a simple construction giving the joint distribution of $\pi_1, \pi_2, \ldots$ and $\theta_1^*, \theta_2^*, \ldots$ called the *stick-breaking construction*.

$$\theta_k^* \sim H$$

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

$$v_k \sim \text{Beta}(1, \alpha)$$



- Also known as the *GEM* distribution, write $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$.

[Sethuraman 1994]

Pólya Urn Scheme

$$\theta_1, \theta_2, \ldots \sim G$$

- The marginal distribution of $\theta_1, \theta_2, \ldots$ has a simple generative process called the *Pólya urn scheme*.

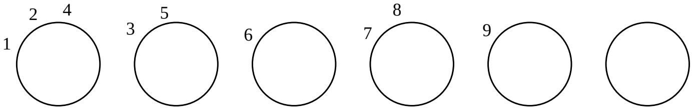
$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- 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 color θ_n into urn.
  - With probability $\propto n - 1$, pick a ball at random from the urn, record θ_n to be its color and return two balls of color θ_n into urn.
- Pólya urn scheme is like a “representer” for the DP—a finite projection of an infinite object G.
- Also known as the *Blackwell-MacQueen urn scheme*.

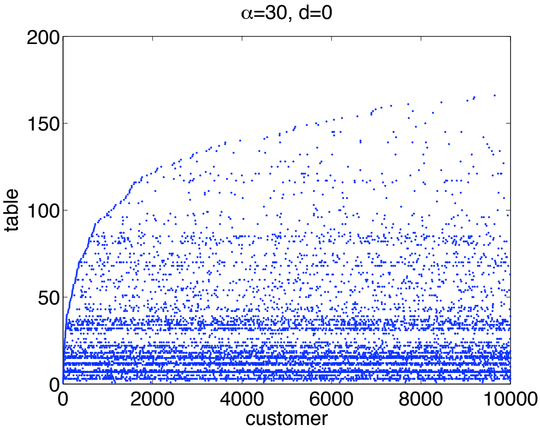
[Blackwell and MacQueen 1973]

Chinese Restaurant Process

- $\theta_1, \ldots, \theta_n$ 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 $\theta_i = \theta_k^*$.
- The distribution over partitions is a *Chinese restaurant process* (CRP).
- 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.



Chinese Restaurant Process



- The CRP exhibits the *clustering property* of the DP.
  - *Rich-gets-richer* effect implies small number of large clusters.
  - Expected number of clusters is $K = O(\alpha \log n)$.

Posterior of Dirichlet Processes

- Since G is a probability measure, we can draw samples from it,

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta_1, \ldots, \theta_n | G &\sim G \end{aligned}$$

  What is the posterior of G given observations of $\theta_1, \ldots, \theta_n$?

- The usual Dirichlet-multinomial conjugacy carries over to the nonparametric DP as well:

$$G | \theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right)$$

Exchangeability

- Instead of deriving the Pólya urn scheme by marginalizing out a DP, consider starting directly from the conditional distributions:

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

- For any n, the joint distribution of $\theta_1, \ldots, \theta_n$ is:

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

  where $h(\theta)$ is density of θ under H, $\theta_1^*, \ldots, \theta_K^*$ are the unique values, and θ_k^* occurred m_{nk} times among $\theta_1, \ldots, \theta_n$.
- The joint distribution is *exchangeable* wrt permutations of $\theta_1, \ldots, \theta_n$.
- *De Finetti's Theorem* says that there must be a random probability measure G making $\theta_1, \theta_2, \ldots$ iid. This is the DP.

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:

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

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

- θ_i's are *independent* given G.
- Sufficient to define G 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, group, arrays,...

Outline

Binary Latent Variable Models

- Consider a latent variable model with binary sources/features,

$$z_{ik} = \begin{cases} 1 & \text{with probability } \mu_k; \\ 0 & \text{with probability } 1 - \mu_k. \end{cases}$$

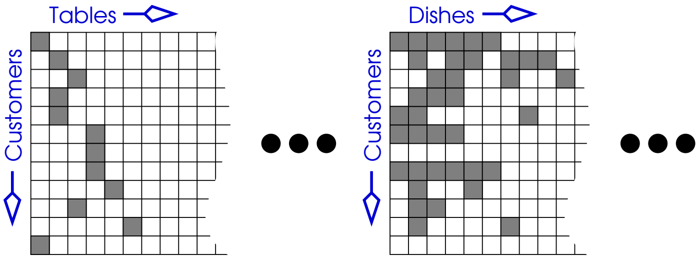
- Example: Data items could be movies like “Terminator 2”, “Shrek” and “Lord of the Rings”, and features could be “science fiction”, “fantasy”, “action” and “Arnold Schwarzenegger”.
- Place beta prior over the probabilities of features:

$$\mu_k \sim \text{Beta}(\tfrac{\alpha}{K}, 1)$$

- We will again take $K \to \infty$.

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 $\frac{m_k}{i}$; and picks Poisson($\frac{\alpha}{i}$) new dishes.



Indian Buffet Processes and Exchangeability

- The IBP is infinitely exchangeable. For this to make sense, we need to “forget” the ordering of the dishes.
  - “Name” each dish k with a Λ^*_k drawn iid from H.
  - Each customer now eats a set of dishes: $\Psi_i = \{\Lambda_k : z_{ik} = 1\}$.
  - The joint probability of $\Psi_1, \ldots, \Psi_n$ can be calculated:

$$p(\Psi_1, \ldots, \Psi_n) = \exp\left(-\alpha \sum_{i=1}^{n} \frac{1}{i}\right) \alpha^K \prod_{k=1}^{K} \frac{(m_k - 1)!(n - m_k)!}{n!} h(\Lambda^*_k)$$

  K: total number of dishes tried by n customers.
  Λ^*_k: Name of kth dish tried.
  m_k: number of customers who tried dish Λ^*_k.

- 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 rate measure:

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

- 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, Ghahramani et al. 2007]

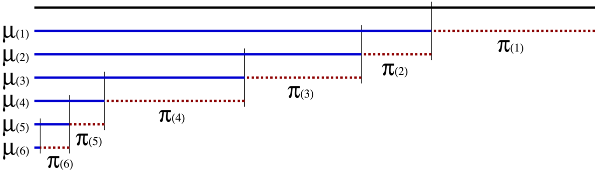
Stick-breaking Construction for Beta Processes

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

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

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

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



[Teh et al. 2007]

Applications of Indian Buffet Processes

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

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

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

- 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].

Infinite 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.

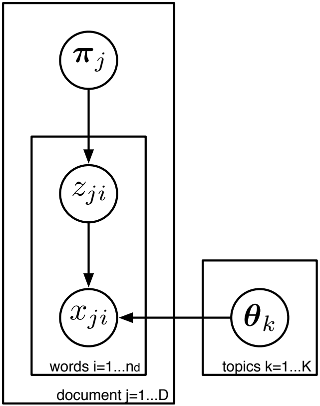
[Knowles and Ghahramani 2007]

Outline

Topic Modelling with Latent Dirichlet Allocation

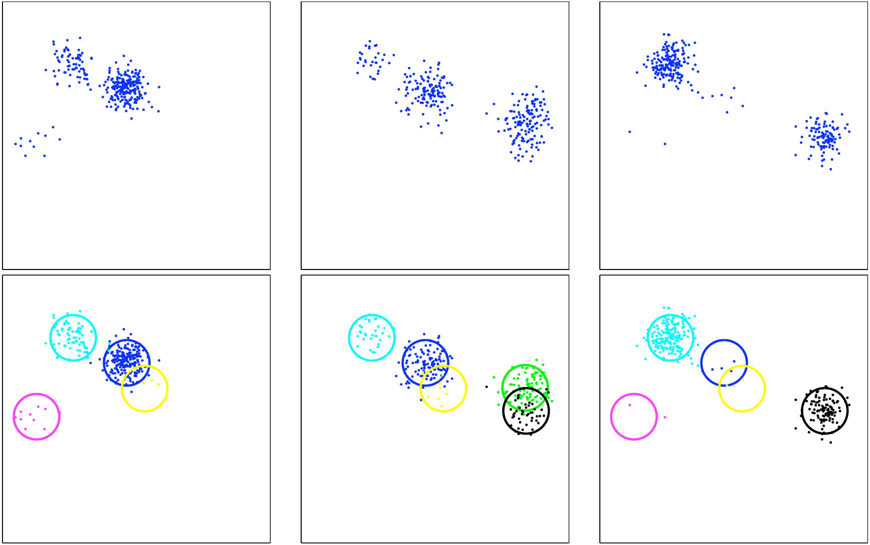
- Infer topics from a document corpus, topics being sets of words that tend to co-occur together.
- Using (Bayesian) latent Dirichlet allocation:

$$
\begin{aligned}
\boldsymbol{\pi}_j &\sim \text{Dirichlet}(\tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\tfrac{\beta}{W}, \ldots, \tfrac{\beta}{W}) \\
z_{ji} | \boldsymbol{\pi}_j &\sim \text{Multinomial}(\boldsymbol{\pi}_j) \\
x_{ji} | z_{ji}, \boldsymbol{\theta}_{z_{ji}} &\sim \text{Multinomial}(\boldsymbol{\theta}_{z_{ji}})
\end{aligned}
$$



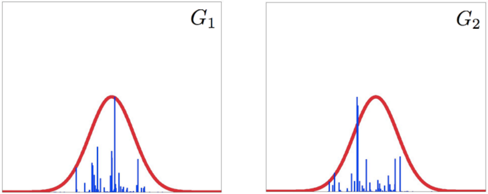
- Can we take $K \to \infty$?

Hierarchical Dirichlet Processes



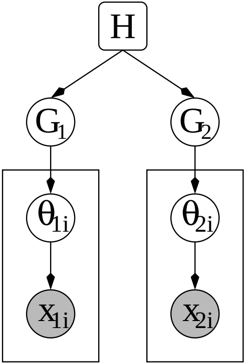
Hierarchical Dirichlet Processes

- 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]

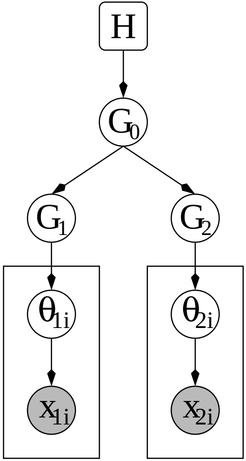


Hierarchical Dirichlet Processes

- A hierarchical Dirichlet process:

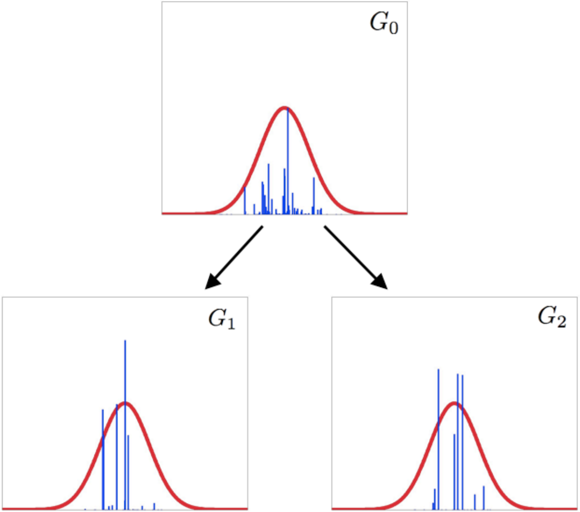
$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0) \text{ iid}$$

- Extension to larger hierarchies is straightforward.



Hierarchical Dirichlet Processes

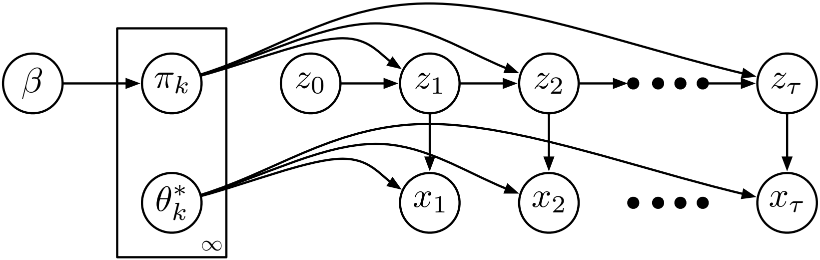
- Making G_0 discrete forces shared cluster between G_1 and G_2.



Hierarchical Dirichlet Processes

- Document topic modelling:
  - Allows documents to be modelled with DP mixtures of topics, with topics shared across corpora.
- Infinite hidden Markov modelling:
  - Allows HMMs with an infinite number of states, with transitions from each allowable state to every other allowable state.
- Learning discrete structures from data:
  - Determining number of objects, nonterminals, states etc.

Infinite Hidden Markov Models



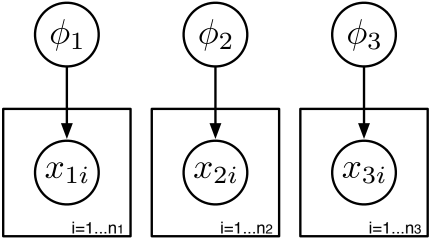
$\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ $\quad$ $\boldsymbol{\pi}_k | \boldsymbol{\beta} \sim \text{DP}(\alpha, \boldsymbol{\beta})$ $\quad$ $z_i | z_{i-1}, \boldsymbol{\pi}_{z_{i-1}} \sim \text{Multinomial}(\boldsymbol{\pi}_{z_{i-1}})$

$\theta_k^* \sim H$ $\quad$ $x_i | z_i, \theta_{z_i}^* \sim F(\theta_{z_i}^*)$

- Hidden Markov models with an infinite number of states.
- Hierarchical DPs used to share information among transition probability vectors prevents “run-away” states.

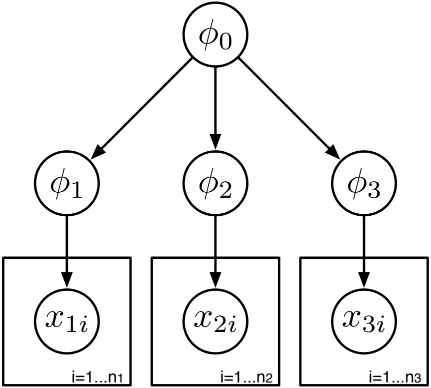
[Beal et al. 2002, Teh et al. 2006]

Hierarchical Modelling



- Better estimation of parameters.
- Multitask learning, learning to learn: generalizing across related tasks.

Hierarchical Modelling



- Better estimation of parameters.
- Multitask learning, learning to learn: generalizing across related tasks.

Outline

Pitman-Yor Processes

- Two-parameter generalization of the Chinese restaurant process:

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

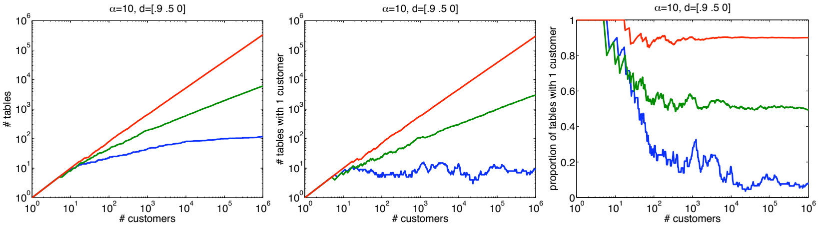
- Associating each cluster k with a unique draw $\theta_k^* \sim H$, the corresponding Pólya urn scheme is also exchangeable.
- De Finetti's Theorem states that there is a random measure underlying this two-parameter generalization.
  - This is the *Pitman-Yor process*.
- The Pitman-Yor process also has a stick-breaking construction:

$$\pi_k = v_k \prod_{i=1}^{k-1}(1 - v_i) \quad \beta_k \sim \text{Beta}(1-\beta, \alpha + \beta k) \quad \theta_k^* \sim H \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

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

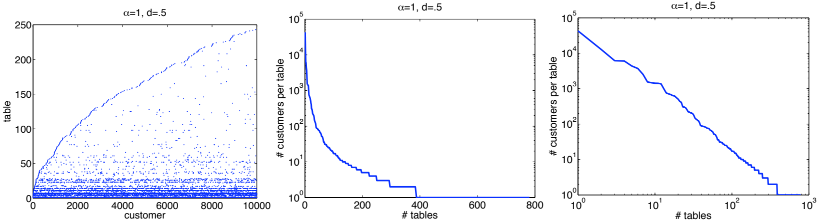
Pitman-Yor Processes

- 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 $K = O(\alpha n^{\beta})$.

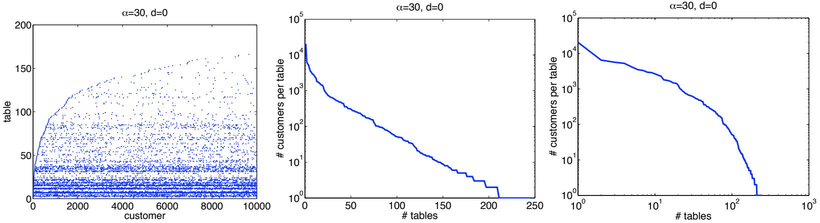


Pitman-Yor Processes

Draw from a Pitman-Yor process



Draw from a Dirichlet process



Hierarchical Pitman-Yor Language Models

- Pitman-Yor processes can be suitable models for many natural phenomena with power-law statistics.
- Language modelling with Markov assumption:

$$\begin{aligned}&p(\text{Mary has a little lamb})\\ \approx\,&p(\text{Mary})p(\text{has}|\text{Mary})p(\text{a}|\text{Mary has})p(\text{little}|\text{has a})p(\text{lamb}|\text{a little})\end{aligned}$$

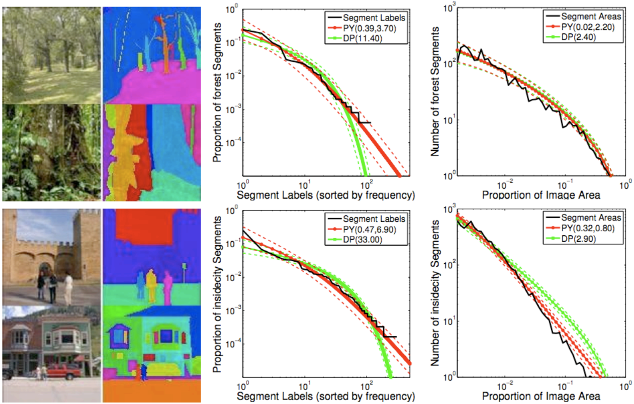
- Parameterize with $p(w_3|w_1,w_2) = G_{w_1,w_2}[w_3]$ and use a hierarchical Pitman-Yor process prior:

$$\begin{aligned}G_{w_1,w_2}|G_{w_2} &\sim \mathrm{PY}(\alpha_2,\beta_2,G_{w_2})\\ G_{w_2}|G_\emptyset &\sim \mathrm{PY}(\alpha_1,\beta_1,G_\emptyset)\\ G_\emptyset|U &\sim \mathrm{PY}(\alpha_0,\beta_0,U)\end{aligned}$$

- State-of-the-art results, connection to Kneser-Ney smoothing.

[Goldwater et al. 2006a, Teh 2006b]

Image Segmentation with Pitman-Yor Processes



- Human segmentations of images also seem to follow power-law.
- An unsupervised image segmentation model based on dependent hierarchical Pitman-Yor processes achieves state-of-the-art results.

[Sudderth and Jordan 2009]

Outline

Summary

- Motivation for Bayesian nonparametrics:
  - Allows practitioners to define and work with models with large support, sidesteps model selection.
  - New models with useful properties.
  - Large variety of applications.
- Various standard Bayesian nonparametric models:
  - Dirichlet processes
  - Hierarchical Dirichlet processes
  - Infinite hidden Markov models
  - Indian buffet and beta processes
  - Pitman-Yor processes
- Touched upon two important theoretical tools:
  - Consistency and Kolmogorov's Consistency Theorem
  - Exchangeability and de Finetti's Theorem
- Described a number of applications of Bayesian nonparametrics.
- Missing: Inference methods based on MCMC, variational etc, consistency and convergence.

Other Introductions to Bayesian Nonparametrics

- Zoubin Gharamani, UAI 2005 Tutorial.
- Michael Jordan, NIPS 2005 Tutorial.
- Volker Tresp, ICML nonparametric Bayes workshop 2006.
- Peter Orbanz, Foundations of Nonparametric Bayesian Methods, 2009.
- I have given a number myself (check webpage).
- I have an introduction to Dirichlet processes [Teh 2007], and another to hierarchical Bayesian nonparametric models [Teh and Jordan 2009].

Bayesian Nonparametric Software

- Hierarchical Bayesian Compiler (HBC). Hal Daume III.
  http://www.cs.utah.edu/ hal/HBC/
- DPpackage. Alejandro Jara.
  http://cran.r-project.org/web/packages/DPpackage/index.html
- Hierarchical Pitman Yor Language Model. Songfang Huang.
  http://homepages.inf.ed.ac.uk/s0562315/progs/index.html
- Nonparametric Bayesian Mixture Models. Yee Whye Teh.
  http://www.gatsby.ucl.ac.uk/ ywteh/research/software.html
- Others...

Outline

Representations of Dirichlet Processes

- Posterior Dirichlet process:

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

- 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:

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

Posterior Dirichlet Processes

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

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

- We are interested in:
  - The marginal distribution of θ with G integrated out.
  - The posterior distribution of G conditioning on θ.

Posterior Dirichlet Processes

Conjugacy between Dirichlet Distribution and Multinomial.

- Consider:

$$\begin{aligned}(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ z|(\pi_1, \ldots, \pi_K) &\sim \text{Discrete}(\pi_1, \ldots, \pi_K)\end{aligned}$$

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

- Then:

$$\begin{aligned}z &\sim \text{Discrete}\left(\tfrac{\alpha_1}{\sum_i \alpha_i}, \ldots, \tfrac{\alpha_K}{\sum_i \alpha_i}\right) \\ (\pi_1, \ldots, \pi_K)|z &\sim \text{Dirichlet}(\alpha_1 + \delta_1(z), \ldots, \alpha_K + \delta_K(z))\end{aligned}$$

  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

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

- Using Dirichlet-multinomial conjugacy,

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

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

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

  i.e. $\theta \sim H$ with G integrated out.

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

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

Posterior Dirichlet Processes

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

Pólya Urn Scheme

- First sample:

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

- Second sample:

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

- n^{th} sample

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

Stick-breaking Construction

- Returning to the posterior process:

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

- 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:

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

- 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:

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

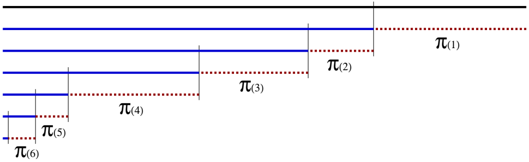
Stick-breaking Construction

- 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

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



Outline

Stick-breaking Construction

- We shall assume the following HDP hierarchy:

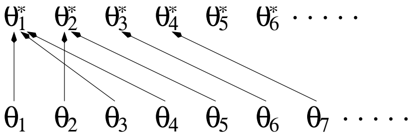
$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_j | G_0 &\sim \mathrm{DP}(\alpha, G_0) \quad \text{for } j = 1, \dots, J \end{aligned}$$

- 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:



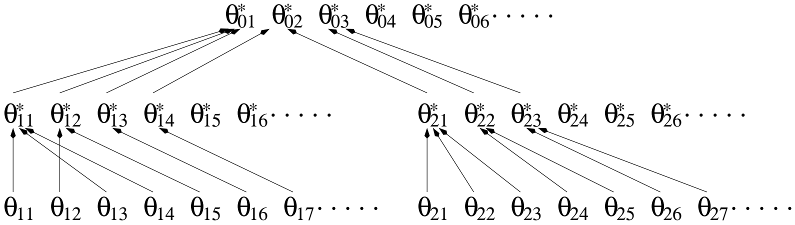
where the arrows denote to which θ^*_k each θ_i was assigned and

$$\theta_1, \theta_2, \ldots \sim G \text{ i.i.d.}$$
$$\theta^*_1, \theta^*_2, \ldots \sim H \text{ i.i.d.}$$

(but $\theta_1, \theta_2, \ldots$ are not independent of $\theta^*_1, \theta^*_2, \ldots$).

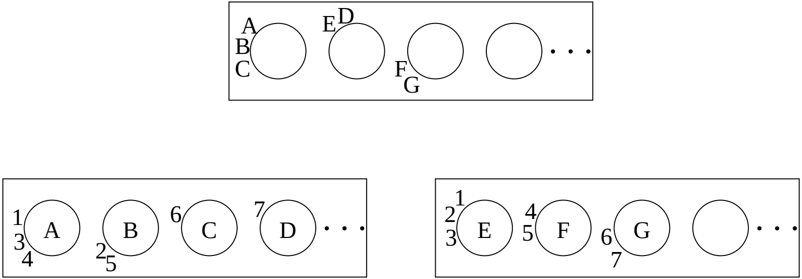
Hierarchical Pòlya Urn Scheme

- 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

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 [Ghahramani et al. 2007]. [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. 2008], 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. 2007a, Kivinen et al. 2007b].
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. 2007, Xing et al. 2006, 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. 2007]. 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. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (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*.

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

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*.

References IV

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

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. (2007).
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.

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. (2007a).
Image denoising with nonparametric hidden Markov trees.
In *IEEE International Conference on Image Processing (ICIP)*, San Antonio, TX.

Kivinen, J., Sudderth, E., and Jordan, M. I. (2007b).
Learning multiscale representations of natural scenes using Dirichlet processes.
In *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

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

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

References VII

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

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.

References VIII

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.

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.

References IX

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.

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.

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. M., 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. and Jordan, M. I. (2009).
Shared segmentation of natural scenes using dependent Pitman-Yor processes.
In *Advances in Neural Information Processing Systems*, volume 21.

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. (2008).
Describing visual scenes using transformed objects and parts.
International Journal of Computer Vision, 77.

References XI

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

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. (2007).
Dirichlet processes.
Submitted to Encyclopedia of Machine Learning.

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. and Jordan, M. I. (2009).
Hierarchical Bayesian nonparametric models with applications.
In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *To appear in Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press.

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.

References XII

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.

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 Sharan, R. (2007).
Bayesian haplotype inference via the Dirichlet process.
Journal of Computational Biology, 14:267–284.

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

References XIII

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.