Bayesian Nonparametrics

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

Acknowledgements:
Cedric Archambeau, Charles Blundell, Hal Daume III, Lloyd Elliott, Jan Gasthaus, Zoubin Ghahramani, Dilan Görür, Katherine Heller, Lancelot James, Michael I. Jordan, Vinayak Rao, Daniel Roy, Jurgen Van Gael, Max Welling, Frank Wood

August, 2010 / CIMAT, Mexico

Outline

Outline

Probabilistic Machine Learning

- *Probabilistic model* of data $\{x_i\}_{i=1}^n$ given parameters θ:

$$P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n|\theta)$$

  where y_i is a latent variable associated with x_i.
- Often thought of as *generative models* of data.
- *Inference*, of latent variables given observations:

$$P(y_1, y_2, \ldots, y_n|\theta, x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n|\theta)}{P(x_1, x_2, \ldots, x_n|\theta)}$$

- *Learning*, typically by *maximum likelihood*:

$$\theta^{\mathrm{ML}} = \operatorname*{argmax}_\theta P(x_1, x_2, \ldots, x_n|\theta)$$

Bayesian Machine Learning

- Probabilistic model of data $\{x_i\}_{i=1}^n$ given parameters θ:

$$P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n|\theta)$$

- *Prior* distribution:

$$P(\theta)$$

- *Posterior* distribution:

$$P(\theta, y_1, y_2, \ldots, y_n|x_1, x_2, \ldots, x_n) = \frac{P(\theta)P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n|\theta)}{P(x_1, x_2, \ldots, x_n)}$$

- *Prediction*:

$$P(x_{n+1}|x_1, \ldots, x_n) = \int P(x_{n+1}|\theta)P(\theta|x_1, \ldots, x_n)d\theta$$

- *(Easier said than done...)*

Computing Posterior Distributions

- Posterior distribution:

$$P(\theta, y_1, y_2, \ldots, y_n | x_1, x_2, \ldots, x_n) = \frac{P(\theta) P(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n | \theta)}{P(x_1, x_2, \ldots, x_n)}$$

- High-dimensional, no closed-form, multi-modal...
- *Variational approximations* [Wainwright and Jordan 2008]: simple parametrized form, "fit" to true posterior.
- *Monte Carlo methods*, including *Markov chain Monte Carlo* [Neal 1993, Robert and Casella 2004] and *sequential Monte Carlo* [Doucet et al. 2001]: construct generators for random samples from the posterior.

Bayesian Model Selection

- *Model selection* is often necessary to prevent *overfitting* and *underfitting*.
- Bayesian approach to model selection uses the *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: $$p(M_k, \theta_k|\mathbf{x}) = \frac{p(M_k)p(\theta_k|M_k)p(\mathbf{x}|\theta_k, M_k)}{\sum_{k'} p(M_{k'})p(\theta_{k'}|M_{k'})p(\mathbf{x}|\theta_{k'}, M_{k'})}$$

- Other approaches to model selection: cross validation, regularization, sparse models...

Side-Stepping Model Selection

- Strategies for model selection often entail significant complexities.
- But reasonable and proper Bayesian methods should not overfit anyway [Rasmussen and Ghahramani 2001].
- Idea: use a large model, and be Bayesian so will not overfit.
- Bayesian nonparametric idea: use a very large Bayesian model avoids both overfitting and underfitting.

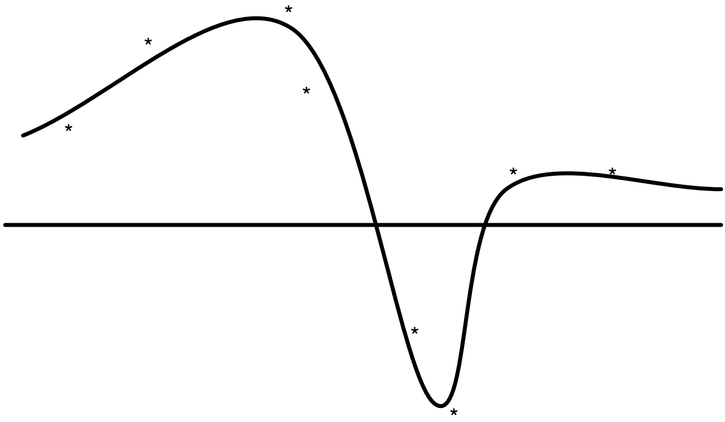
Direct Modelling of Very Large Spaces

- Regression: learn about *functions* from an input to an output space.
- Density estimation: learn about *densities* over $\mathbb{R}^d$.
- Clustering: learn about *partitions* of a large space.
- Objects of interest are often infinite dimensional. Model these directly:
  - Using models that can learn any such object;
  - Using models that can approximate any such object to arbitrary accuracy.
- Many theoretical and practical issues to resolve:
  - Convergence and consistency.
  - Practical inference algorithms.

Outline

Regression and Classification

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



- Regression: if $y_i = f^*(x_i) + \epsilon_i$.
- Classification: e.g. $P(y_i = 1 | f^*(x_i)) = \Phi(f^*(x_i))$.

Parametric 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

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

- What family of basis function to use?
- How many?
- What if true function cannot be parametrized as such?

Towards Nonparametric Regression

- What we are interested in is the output values of the function,

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

  Why not model these directly?

- In regression, each $f(x_i)$ is continuous and real-valued, so a natural choice is to model $f(x_i)$ using a Gaussian.
- Assume that function f is *smooth*. If two inputs x_i and x_j are close-by, then $f(x_i)$ and $f(x_j)$ should be close by as well. This translates into *correlations* among the outputs $f(x_i)$.

Towards Nonparametric Regression

- We can use a multi-dimensional Gaussian to model correlated function outputs:

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_{n+1}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} C_{1,1} & \dots & C_{1,n+1} \\ \vdots & \ddots & \vdots \\ C_{n+1,1} & \dots & C_{n+1,n+1} \end{bmatrix} \right)$$

  where the mean is zero, and $C = [C_{ij}]$ is the covariance matrix.
- Each observed output y_i can be modelled as,

$$y_i | f(x_i) \sim \mathcal{N}(f(x_i), \sigma^2)$$

- Learning: compute posterior distribution

$$p(f(x_1), \dots, f(x_n) | y_1, \dots, y_n)$$

  Straightforward since whole model is Gaussian.
- Prediction: compute

$$p(f(x_{n+1}) | y_1, \dots, y_n)$$

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 $c(x, y)$.
- Difference from before: the GP defines a distribution over $f(x)$, for *every* input value x simultaneously. Prior is defined even before observing inputs $x_1, \ldots, x_n$.
- Such a random function f is known as a *stochastic process*. It is a collection of random variables $\{f(x)\}_{x \in \mathbb{X}}$.
- Demo: GPgenerate.

[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.

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

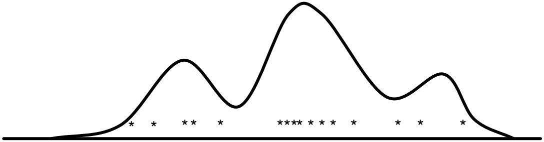
    $$p_{\mathbf{x}}([f(x)]_{x\in\mathbf{x}}) = \int p_{\mathbf{x}\cup\mathbf{y}}([f(x)]_{x\in\mathbf{x}\cup\mathbf{y}}) d[f(x)]_{x\in\mathbf{y}}$$

    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 $\{f(x)\}_{x\in\mathbb{X}}$.

Outline

Density Estimation with Mixture Models

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



- Can use a mixture model for flexible family of densities, e.g.

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

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

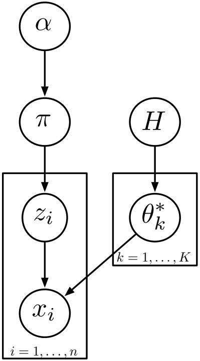
Bayesian Mixture Models

- Let's be Bayesian about mixture models, and place priors over our parameters (and to compute posteriors).
- First, introduce conjugate priors for parameters:

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

- Second, introduce variable z_i indicator which component x_i belongs to.

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

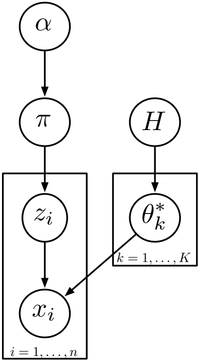


[Rasmussen 2000]

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:

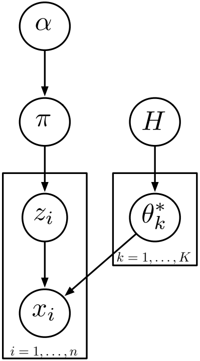
  Occupied clusters:

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

  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

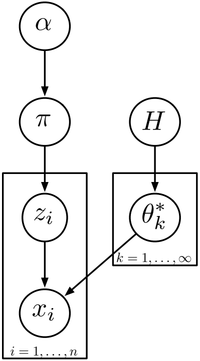
- We will take $K \to \infty$.
- Imagine a very large value of K.
- There are at most $n < K$ occupied components, so most components are *empty*. We can lump these empty components together:

  Occupied clusters:

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

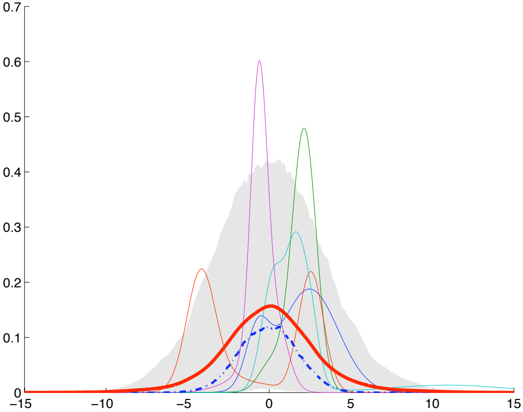
  Empty clusters:

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



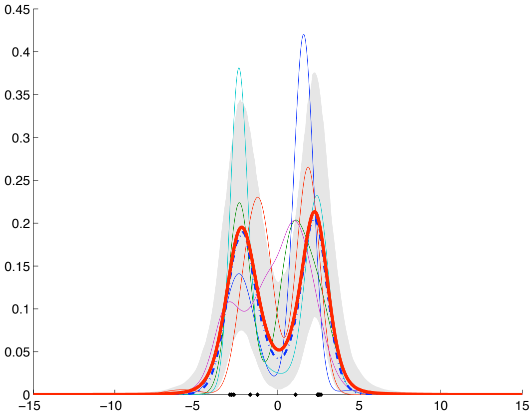
- Demo: dpm_demointeractive.

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: posterior samples. 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: posterior samples. Black: data points.

Infinite Bayesian Mixture Models

- The actual infinite limit of finite mixture models does not actually make mathematical sense.
- 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).
- The $K \to \infty$ Gibbs sampler is for DP mixture models.

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.
  - Can think of a stochastic process as a *random function* $X(i)$.
- Stochastic processes form the core of many Bayesian nonparametric models.
  - Gaussian processes, Poisson processes, Dirichlet processes, beta processes, completely random measures...

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,

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

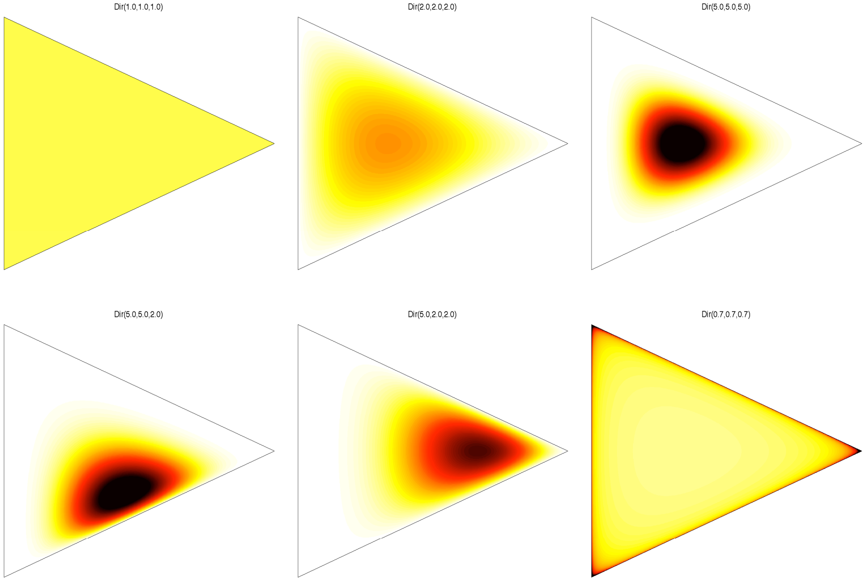
  with parameters $(\lambda_1, \dots, \lambda_K)$, if

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

- Equivalent to normalizing a set of independent gamma variables:

$$\begin{aligned}(\pi_1, \dots, \pi_K) &= \tfrac{1}{\sum_k \gamma_k}(\gamma_1, \dots, \gamma_K) \\ \gamma_k &\sim \text{Gamma}(\lambda_k) \qquad \text{for } k = 1, \dots, K\end{aligned}$$

Dirichlet Distributions

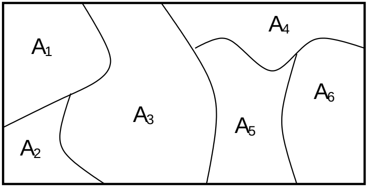


Dirichlet Processes

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

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

where λ is a base measure.



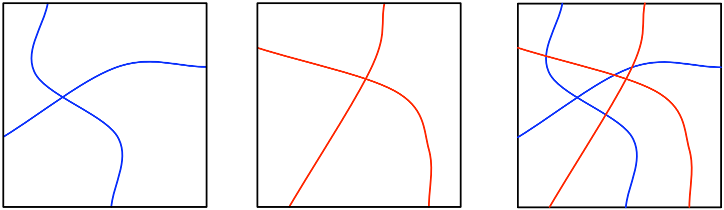
- The above family of distributions is consistent (next slide), and *Kolmogorov Consistency Theorem* can be applied to show existence (but there are technical conditions restricting the generality of the definition).

[Ferguson 1973, Blackwell and MacQueen 1973]

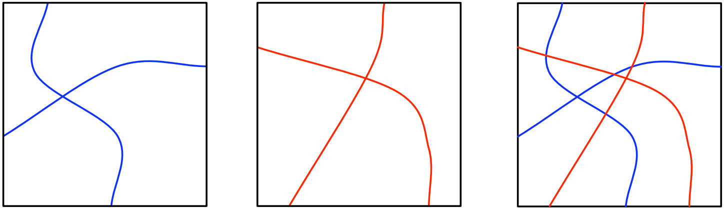
Consistency of Dirichlet Marginals

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

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

so the distributions of $(G(A_1), \dots, G(A_K))$ and $(G(B_1), \dots, G(B_J))$ 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 $(A_1, \ldots, A_K)$ of Θ:

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

- The first and second moments 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.

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.
  - Can we describe G by describing its effect $\theta_1, \theta_2, \ldots$?
  - What is the marginal distribution of $\theta_1, \theta_2, \ldots$ with G integrated out?

Stick-breaking Construction

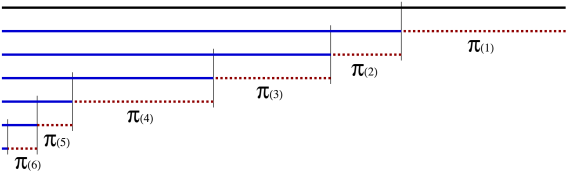
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

- 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]

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:

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

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* (aka *Blackwell-MacQueen urn scheme*).

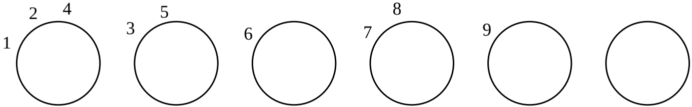
$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- 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.

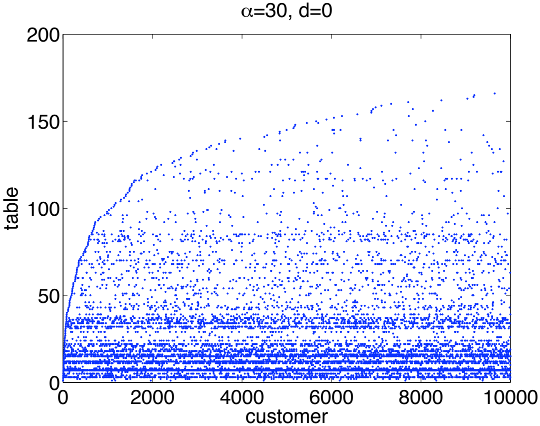
[Blackwell and MacQueen 1973]

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.



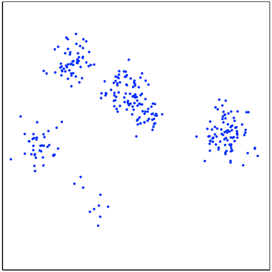
Chinese Restaurant Process



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

Clustering

- To partition a heterogeneous data set into distinct, homogeneous clusters.



- The CRP is a canonical nonparametric prior over partitions that can be used as part of a Bayesian model for clustering.
- Other priors over partitions can be used instead of the CRP induced by a DP (for examples see [Lijoi and Pruenster 2010]).

Inferring Discrete Latent Structures

- DPs have also found uses in applications where the aim is to discover latent objects, and where the number of objects is not known or unbounded.
  - Nonparametric probabilistic context free grammars.
  - Visual scene analysis.
  - Infinite hidden Markov models/trees.
  - Genetic ancestry inference.
  - ...
- In many such applications it is important to be able to model the same set of objects in different contexts.
- This can be tackled using *hierarchical Dirichlet processes*.

[Teh et al. 2006, Teh and Jordan 2010]

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 (unique) 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*).

Outline

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?

Binary Latent Variable Models

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

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

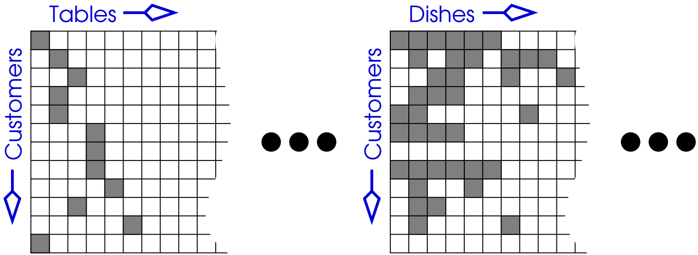
- Example: Data items could be movies like "Terminator 2", "Shrek" and "Lord of the Rings", and features could be "science fiction", "fantasy", "action" and "Arnold Schwarzenegger".
- Place beta prior over the probabilities of features:

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

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

Indian Buffet Processes

- The *Indian Buffet Process* (IBP) 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.



[Griffiths and Ghahramani 2006]

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]

Applications of Indian Buffet Processes

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

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

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

- 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 et al. 2006].
- Independent components analysis [Knowles and Ghahramani 2007].
- Learning the structure of deep belief networks [Adams et al. 2010].

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:

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

- An ICA model with infinite number of features.

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

Beta Processes

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

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

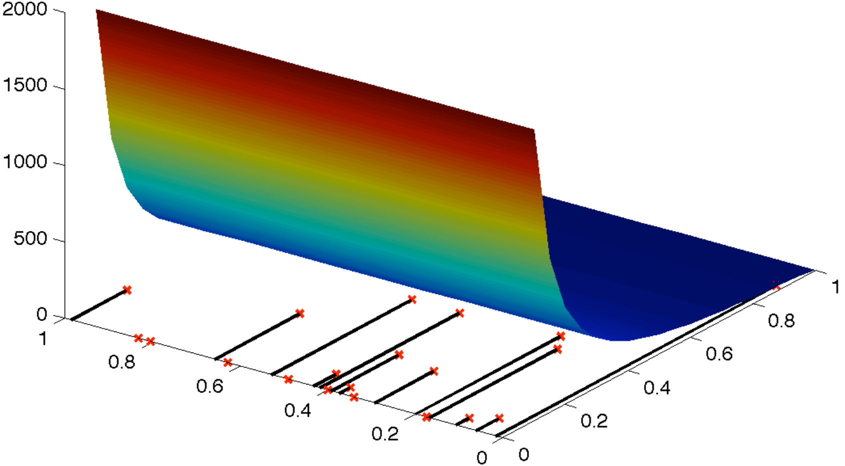
  where the points $P = \{(\theta_1^*, \mu_1), (\theta_2^*, \mu_2), \ldots\}$ are spikes in a 2D Poisson process with rate measure:

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

- It is the de Finetti measure for the IBP.
- This is an example of a *completely random measure*.
- A beta process *does not* have Beta distributed marginals.

[Hjort 1990, Thibaux and Jordan 2007]

Beta Processes



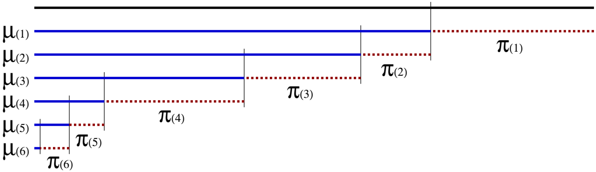
Stick-breaking Construction for Beta Processes

- The following generates a draw of B:

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

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

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



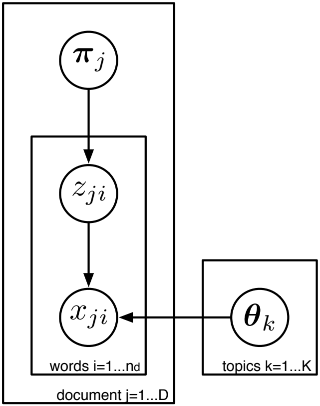
[Teh et al. 2007]

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:

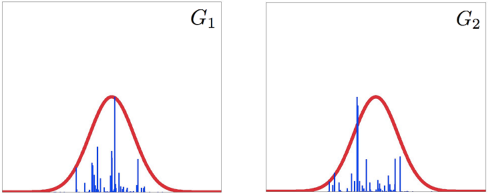
$$\begin{aligned}
\boldsymbol{\pi}_j &\sim \text{Dirichlet}(\tfrac{\alpha}{K}, \dots, \tfrac{\alpha}{K}) \\
\boldsymbol{\theta}_k &\sim \text{Dirichlet}(\tfrac{\beta}{W}, \dots, \tfrac{\beta}{W}) \\
z_{ji} | \boldsymbol{\pi}_j &\sim \text{Multinomial}(\boldsymbol{\pi}_j) \\
x_{ji} | z_{ji}, \boldsymbol{\theta}_{z_{ji}}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

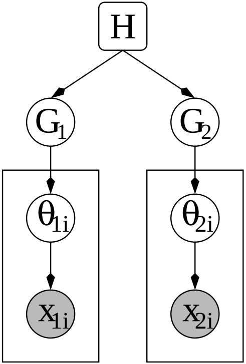


- How many topics can we find from the corpus?
- Can we take number of topics $K \to \infty$?

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.

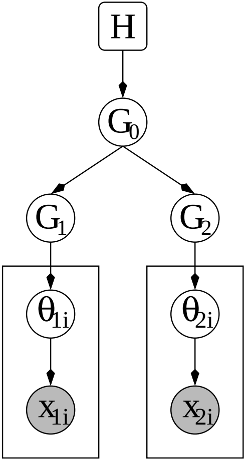
[Teh et al. 2006]

Hierarchical Dirichlet Processes

- A hierarchical Dirichlet process:

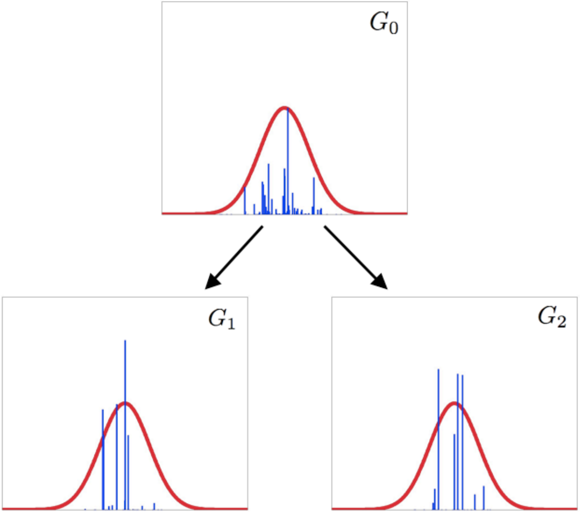
$$G_0 \sim \mathrm{DP}(\alpha_0, H)$$
$$G_1, G_2 | G_0 \sim \mathrm{DP}(\alpha, G_0) \text{ iid}$$

- Extension to larger hierarchies is straightforward.

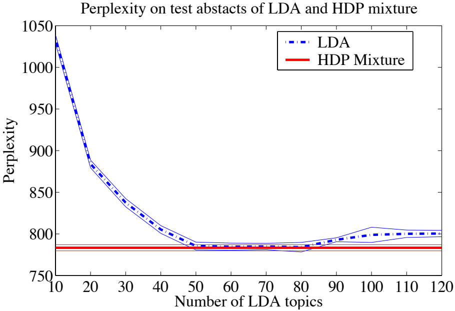


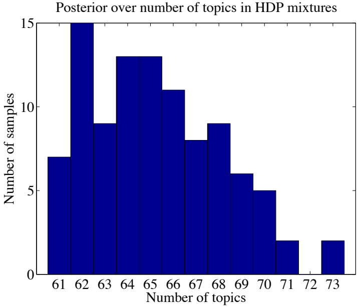
Hierarchical Dirichlet Processes

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

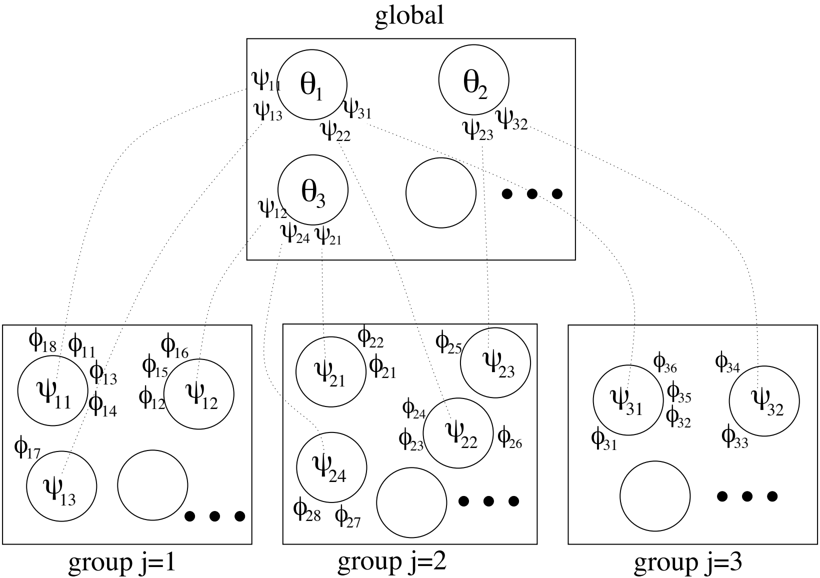


Hierarchical Dirichlet Processes

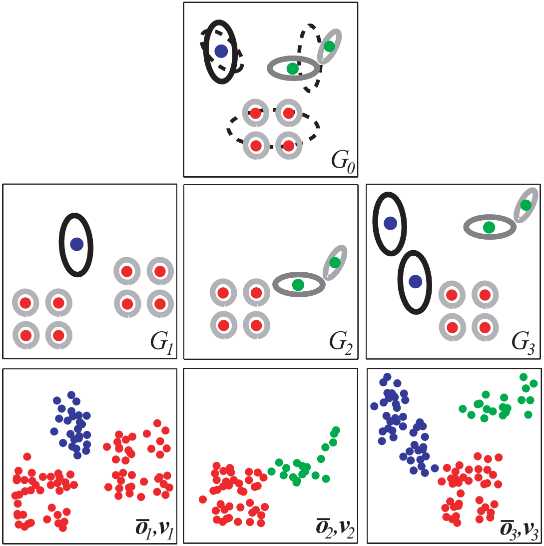




Chinese Restaurant Franchise

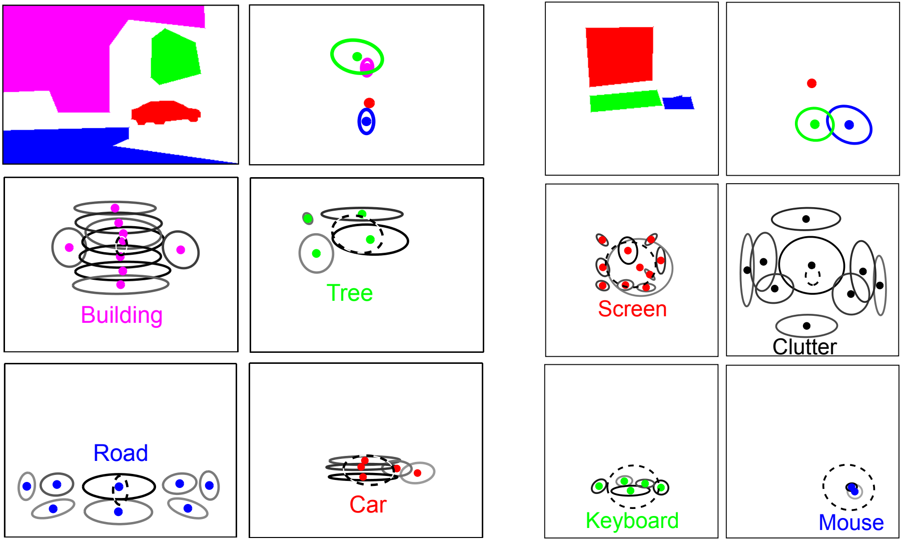


Visual Scene Analysis with Transformed DPs



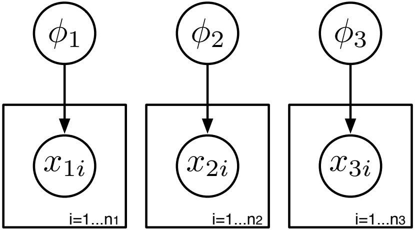
[Sudderth et al. 2008]

Visual Scene Analysis with Transformed DPs



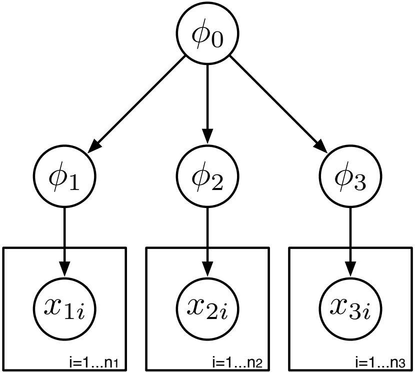
[Sudderth et al. 2008]

Hierarchical Modelling



[Gelman et al. 1995]

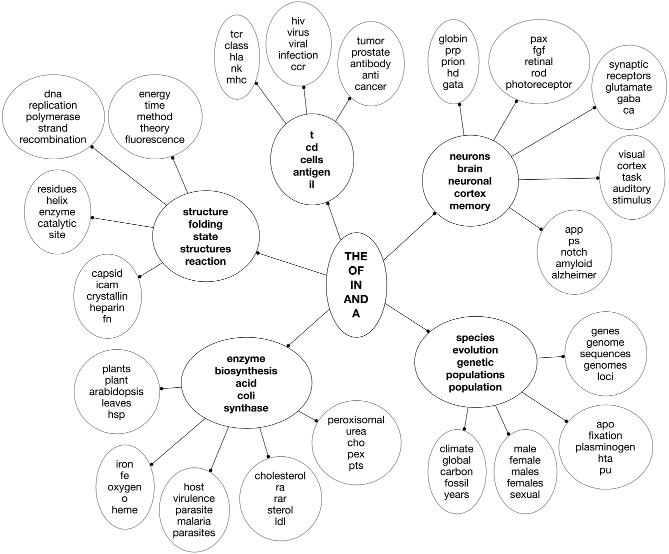
Hierarchical Modelling



[Gelman et al. 1995]

Outline

Topic Hierarchies



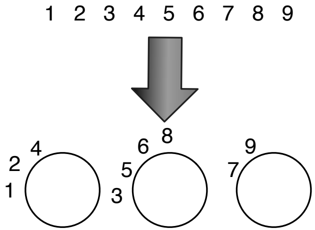
[Blei et al. 2010]

Nested Chinese Restaurant Process

1 2 3 4 5 6 7 8 9

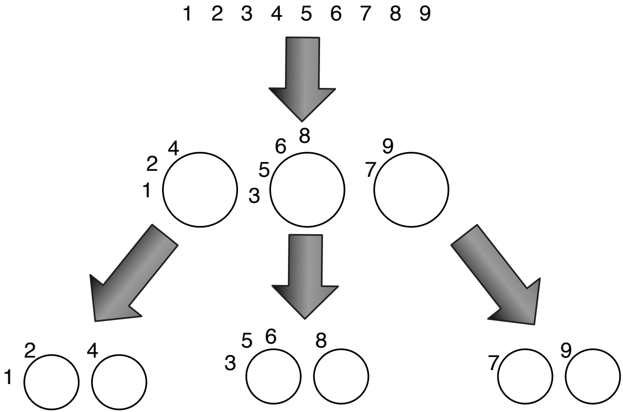
[Blei et al. 2010]

Nested Chinese Restaurant Process



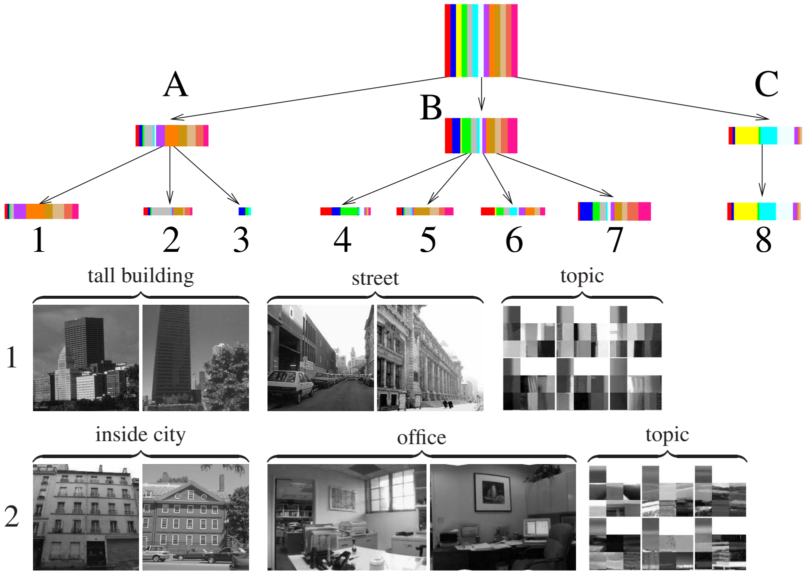
[Blei et al. 2010]

Nested Chinese Restaurant Process



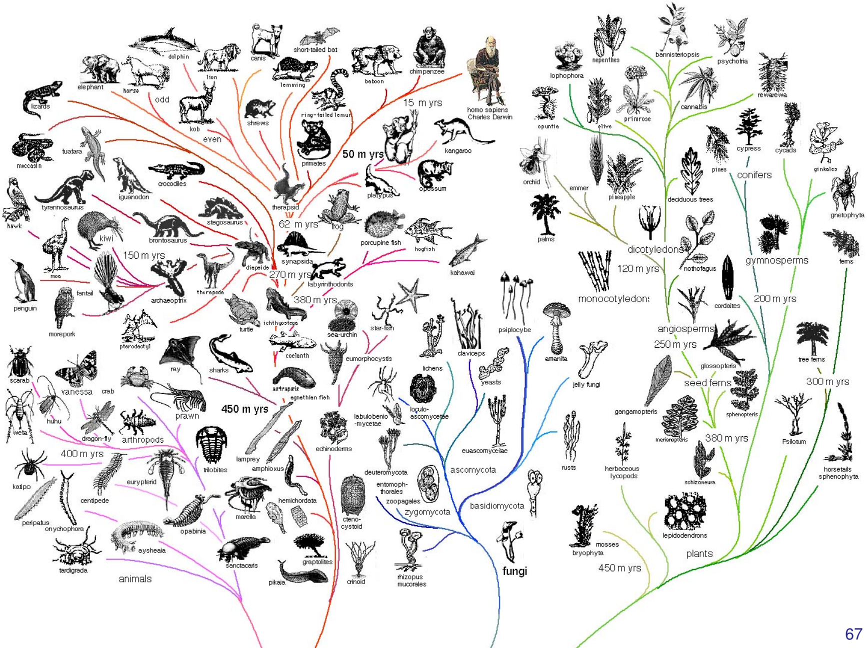
[Blei et al. 2010]

Visual Taxonomies



[Bart et al. 2008]

Hierarchical Clustering



Hierarchical Clustering

- Bayesian approach to hierarchical clustering: place prior over tree structures, and infer posterior.
- The nested DP can be used as a prior over layered tree structures.
- Another prior is a *Dirichlet diffusion tree*, which produces binary ultrametric trees, and which can be obtained as an infinitesimal limit of a nested DP. It is an example of a *fragmentation process*.
- Yet another prior is *Kingman's coalescent*, which also produces binary ultrametric trees, but is an example of a *coalescent process*.

[Neal 2003, Teh et al. 2008, Bertoin 2006]

Nested Dirichlet Process

- Underlying stochastic process for the nested CRP is a *nested DP*.

Hierarchical DP:

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\alpha_0, H) \\
G_j | G_0 &\sim \mathrm{DP}(\alpha, G_0) \\
x_{ji} | G_j &\sim G_j
\end{aligned}
$$

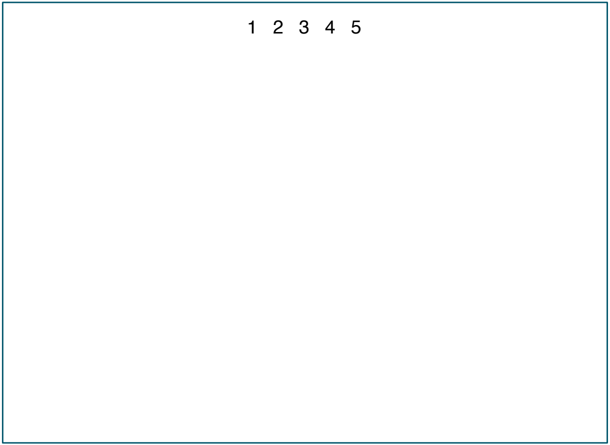
Nested DP:

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

- The hierarchical DP starts with groups of data items, and analyses them together by introducing dependencies through G_0.
- The nested DP starts with one set of data items, partitions them into different groups, and analyses each group separately.
- Orthogonal effects, can be used together.

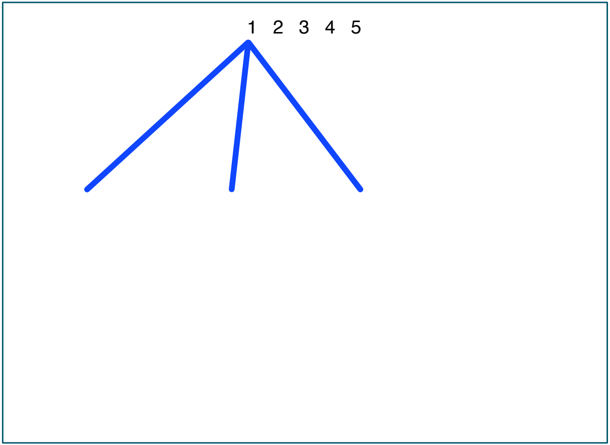
[Rodríguez et al. 2008]

Nested Beta/Indian Buffet Processes



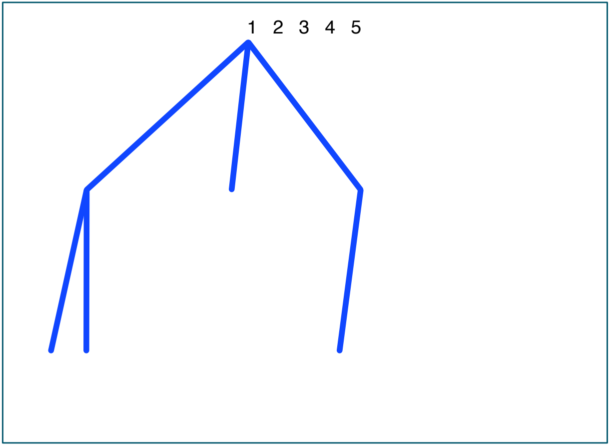
- Exchangeable distribution over *layered trees*.

Nested Beta/Indian Buffet Processes



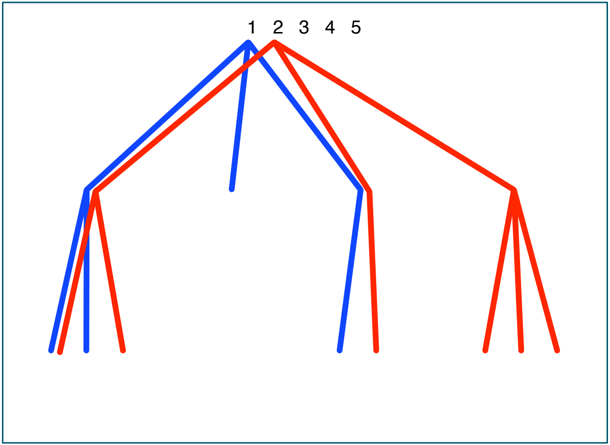
- Exchangeable distribution over *layered trees*.

Nested Beta/Indian Buffet Processes



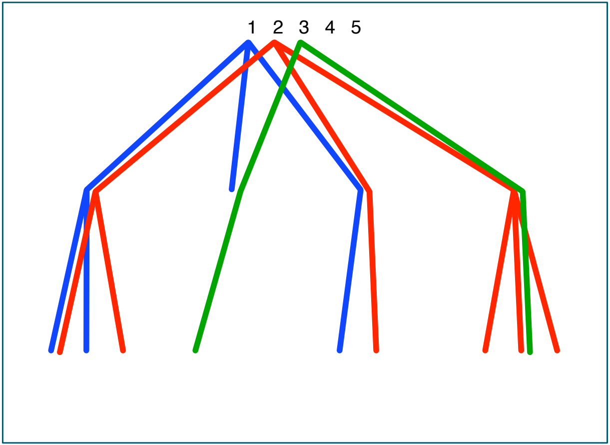
- Exchangeable distribution over *layered trees*.

Nested Beta/Indian Buffet Processes



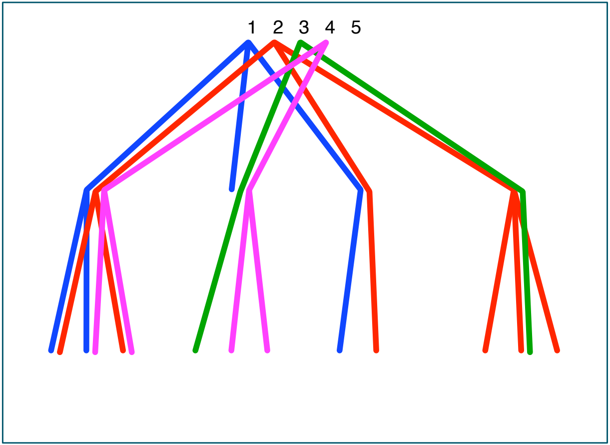
- Exchangeable distribution over *layered trees*.

Nested Beta/Indian Buffet Processes



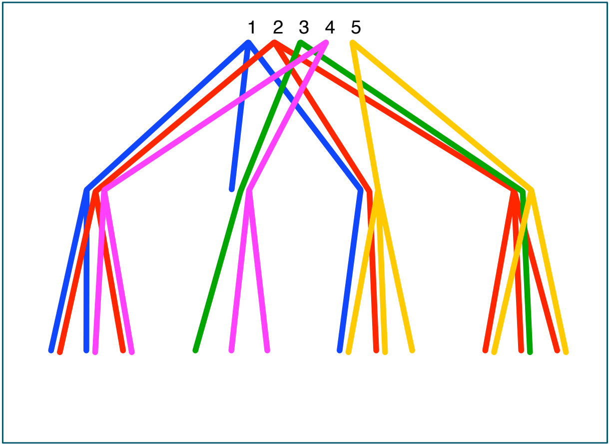
- Exchangeable distribution over *layered trees*.

Nested Beta/Indian Buffet Processes



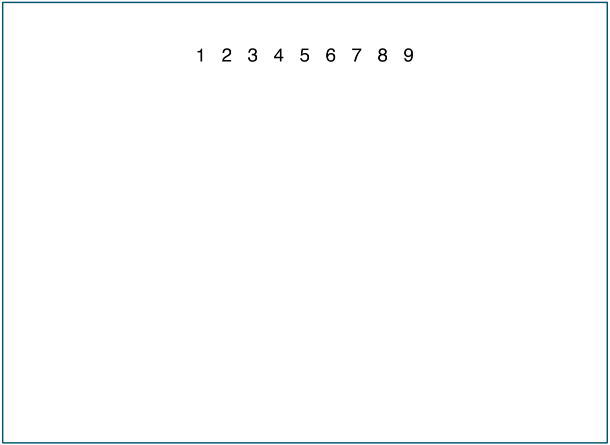
- Exchangeable distribution over *layered trees*.

Nested Beta/Indian Buffet Processes



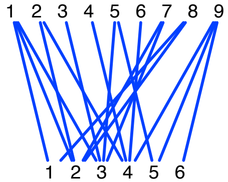
- Exchangeable distribution over *layered trees*.

Hierarchical Beta/Indian Buffet Processes



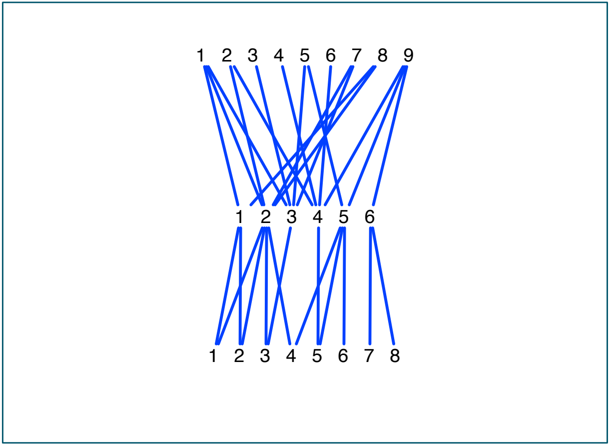
- Different from the *hierarchical beta process* of [Thibaux and Jordan 2007].

Hierarchical Beta/Indian Buffet Processes



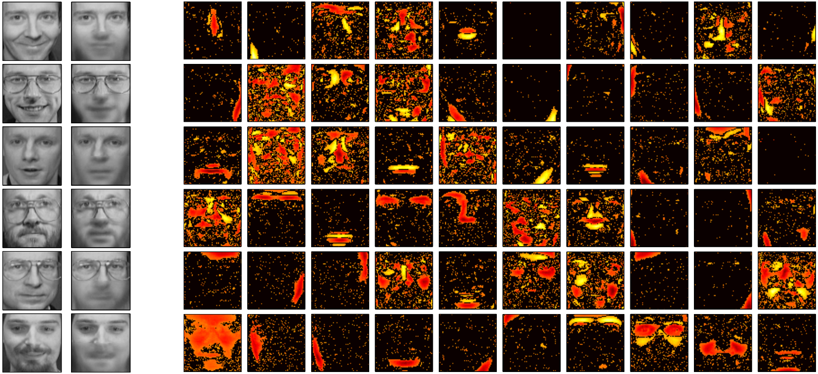
- Different from the *hierarchical beta process* of [Thibaux and Jordan 2007].

Hierarchical Beta/Indian Buffet Processes



- Different from the *hierarchical beta process* of [Thibaux and Jordan 2007].

Deep Structure Learning



[Adams et al. 2010]

Deep Structure Learning

[Adams et al. 2010]

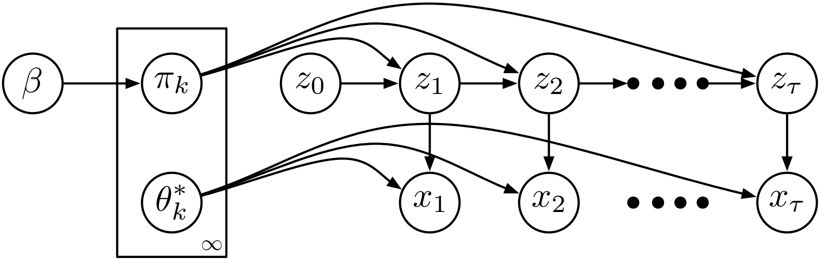
Transfer Learning

- Many recent machine learning paradigms can be understood as trying to model data from heterogeneous sources and types.
  - *Semi-supervised learning*: we have labelled data, and unlabelled data.
  - *Multi-task learning*: we have multiple tasks with different distributions but structurally similar.
  - *Domain adaptation*: we have a small amount of pertinent data, and a large amount of data from a related problem or domain.
- The *transfer learning* problem is how to transfer information between different sources and types.
- Flexible nonparametric models can allow for more information extraction and transfer.
- Hierarchies and nestings are different ways of putting together multiple stochastic processes to form complex models.

[Jordan 2010]

Outline

Hidden Markov Models



$\pi_k \sim \text{Dirichlet}(\frac{\alpha}{K}, \dots, \frac{\alpha}{K})$

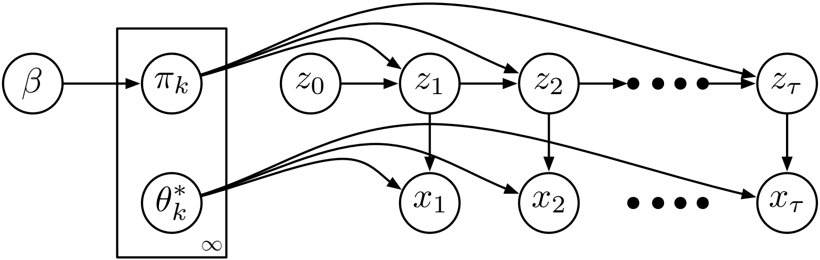
$\theta_k^* \sim H$

$z_i | z_{i-1}, \pi_{z_{i-1}} \sim \text{Multinomial}(\pi_{z_{i-1}})$

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

- Can we take $K \to \infty$?
- Can we do so while imposing structure in transition probability matrix?

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: *infinite HMM*.
- Hierarchical DPs used to share information among transition probability vectors prevents "run-away" states: *HDP-HMM*.

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

Word Segmentation

- Given sequences of utterances or characters can a probabilistic model segment sequences into coherent chunks (“words”)?

canyoureadthissentencewithoutspaces?
can you read this sentence without spaces?
金庸曾把所創作的小說名稱的首字聯成一副對聯：飛雪連天射白鹿，笑書神俠倚碧鴛。

- Use an infinite HMM: each chunk/word is a state, with Markov model of state transitions.
- Nonparametric model is natural, since number of words unknown before segmentation.

[Goldwater et al. 2006b]

Word Segmentation

	Words	Lexicon	Boundaries
NGS-u	68.9	82.6	52.0
MBDP-1	68.2	82.3	52.4
DP	53.8	74.3	57.2
NGS-b	68.3	82.1	55.7
HDP	**76.6**	**87.7**	**63.1**

- NGS-u: *n*-gram Segmentation (unigram) [Venkataraman 2001].
- NGS-b: *n*-gram Segmentation (bigram) [Venkataraman 2001].
- MBDP-1: Model-based Dynamic Programming [Brent 1999].
- DP, HDP: Nonparametric model, without and with Markov dependencies.

[Goldwater et al. 2006a]

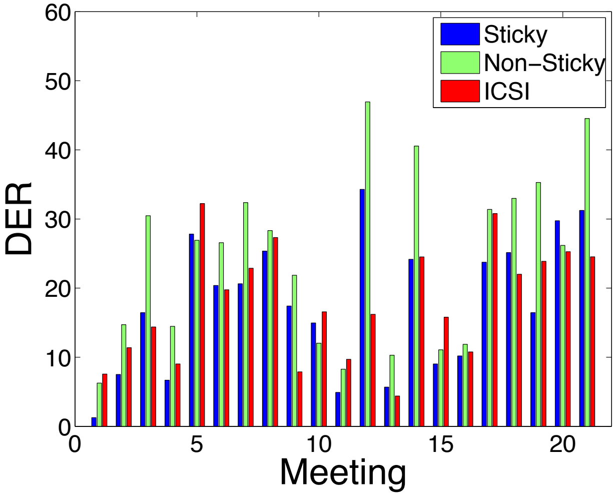
Sticky HDP-HMM

- In typical HMMs or in infinite HMMs the model does not give special treatment to self-transitions (from a state to itself).
- In many HMM applications self-transitions are much more likely.
- Example application of HMMs: speaker diarization.
- Straightforward extension of HDP-HMM prior encourages higher self-transition probabilities:

$$\boldsymbol{\pi}_k|\boldsymbol{\beta} \sim \mathrm{DP}(\alpha+\kappa, \tfrac{\alpha\boldsymbol{\beta}+\kappa\delta_k}{\alpha+\kappa})$$

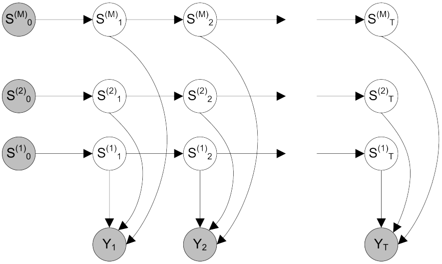
[Beal et al. 2002, Fox et al. 2008]

Sticky HDP-HMM



[Fox et al. 2008]

Infinite Factorial HMM



- Take $M \to \infty$ for the following model specification:

$$P(s_t^{(m)} = 1 | s_{t-1}^{(m)} = 0) = a_m \qquad a_m \sim \text{Beta}(\tfrac{\alpha}{M}, 1)$$

$$P(s_t^{(m)} = 1 | s_{t-1}^{(m)} = 1) = b_m \qquad b_m \sim \text{Beta}(\gamma, \delta)$$

- Stochastic process is a *Markov Indian buffet process*. It is an example of a *dependent random measure*.

[Van Gael et al. 2009]

Nonparametric Grammars, Hierarchical HMMs etc

- In linguistics, grammars are much more plausible as generative models of sentences.
- Learning the structure of probabilistic grammars is even more difficult, and Bayesian nonparametrics provides a compelling alternative.

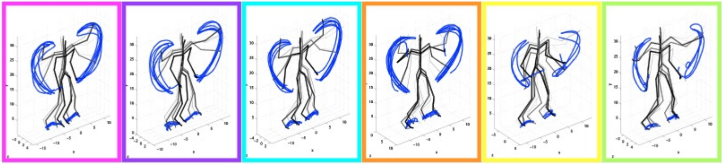
[Liang et al. 2007, Finkel et al. 2007, Johnson et al. 2007, Heller et al. 2009]

Motion Capture Analysis

- Goal: find coherent “behaviour” in the time series that transfers to other time series.

Slides courtesy of [Fox et al. 2010]

Motion Capture Analysis

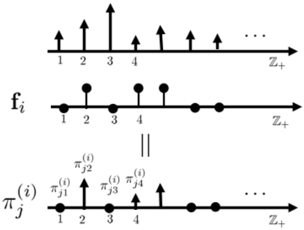


- Transfer knowledge among related time series in the form of a library of "behaviours".
- Allow each time series model to make use of an arbitrary subset of the behaviours.
- Method: represent behaviors as states in an autoregressive HMM, and use the beta/Bernoulli process to pick out subsets of states.

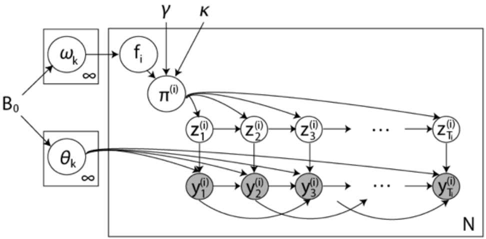
Slides courtesy of [Fox et al. 2010]

BP-AR-HMM

- Bernoulli process determines which states are used



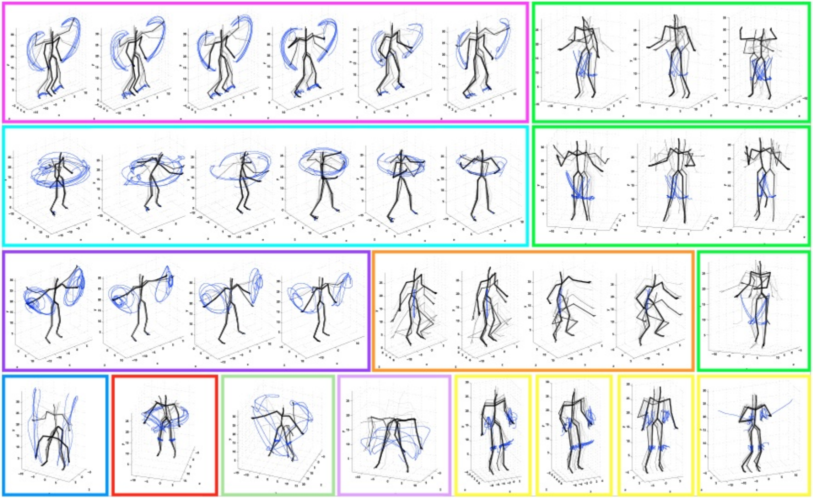
- Beta process prior:
  - encourages sharing
  - allows variability



$$
\begin{aligned}
\pi_j^{(i)} \mid \mathbf{f}_i, \gamma, \kappa &\sim \mathrm{Dir}([\gamma, \ldots, \gamma, \gamma+\kappa, \gamma, \ldots] \otimes \mathbf{f}_i) \\
z_t^{(i)} &\sim \pi_{z_{t-1}^{(i)}}^{(i)} \\
\mathbf{y}_t^{(i)} &= \sum_{j=1}^{r} A_{j, z_t^{(i)}} \mathbf{y}_{t-j}^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)})
\end{aligned}
$$

Slides courtesy of [Fox et al. 2010]

Motion Capture Results



Slides courtesy of [Fox et al. 2010]

High Order Markov Models

- Decompose the joint distribution of a sequence of variables into conditional distributions:

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t | x_1, \ldots, x_{t-1})$$

- An Nth order Markov model approximates the joint distribution as:

$$P(x_1, x_2, \ldots, x_T) = \prod_{t=1}^{T} P(x_t | x_{t-N}, \ldots, x_{t-1})$$

- Such models are particularly prevalent in natural language processing, compression and biological sequence modelling.

toad, in, a, hole
t, o, a, d, _, i, n, _, a, _, h, o, l, e
A, C, G, T, C, C, A

- Would like to take $N \to \infty$.

High Order Markov Models

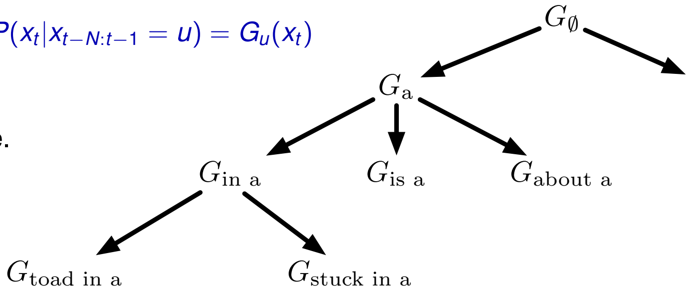
- Difficult to fit such models due to data sparsity.

$$P(x_t|x_{t-N},\ldots,x_{t-1}) = \frac{C(x_{t-N},\ldots,x_{t-1},x_t)}{C(x_{t-N},\ldots,x_{t-1})}$$

- Sharing information via hierarchical models.

$$P(x_t|x_{t-N:t-1} = u) = G_u(x_t)$$

- A context tree.



[MacKay and Peto 1994, Teh 2006a]

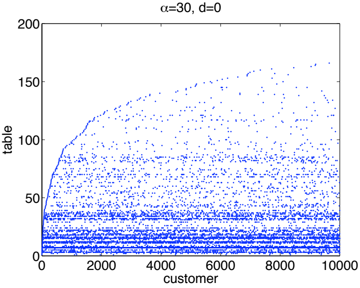
Outline

Pitman-Yor Processes

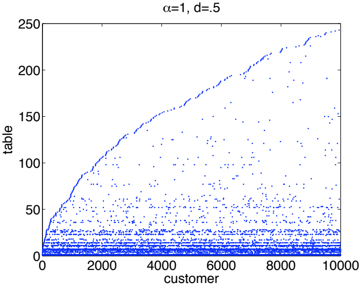
- Two-parameter generalization of the Chinese restaurant process:

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

- Associating each cluster k with a unique draw $\theta_k^* \sim H$, the corresponding Pólya urn scheme is also exchangeable.



Dirichlet



Pitman-Yor

Pitman-Yor Processes

- 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:

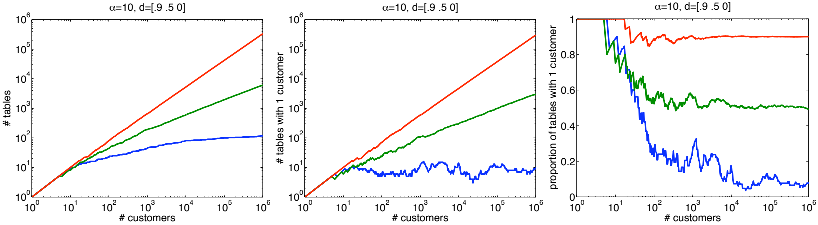
$$\pi_k = v_k \prod_{i=1}^{k-1}(1 - v_i) \quad \beta_k \sim \text{Beta}(1 - \beta, \alpha + \beta k) \quad \theta_k^* \sim H \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

- The Pitman-Yor process cannot be obtained as the infinite limit of a simple parametric model.

[Perman et al. 1992, Pitman and Yor 1997, Ishwaran and James 2001]

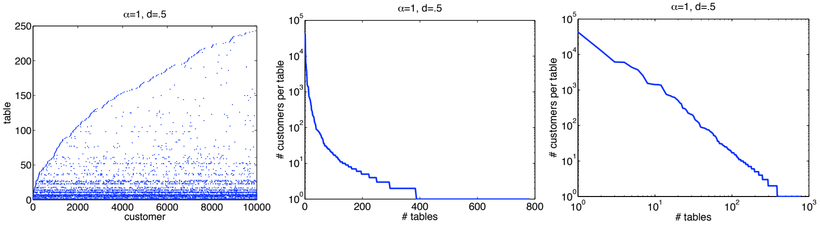
Pitman-Yor Processes

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

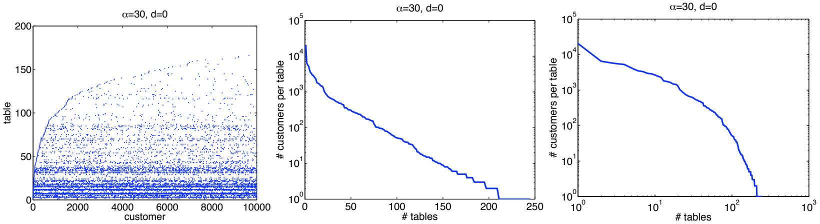


Pitman-Yor Processes

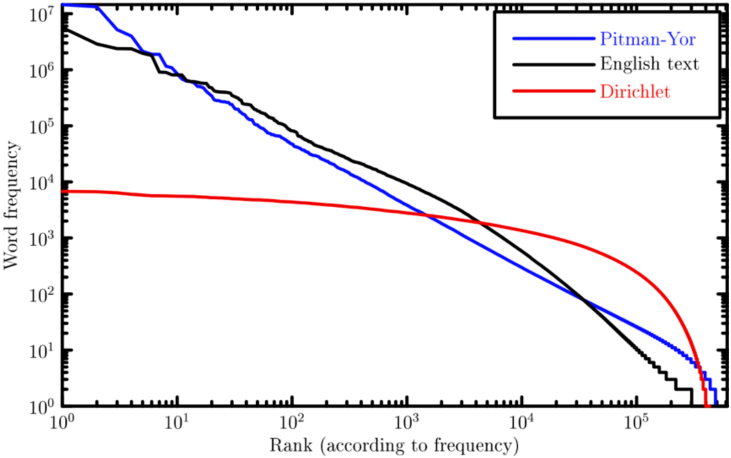
Draw from a Pitman-Yor process



Draw from a Dirichlet process



Pitman-Yor Processes



Hierarchical Pitman-Yor Markov Models

- Use a hierarchical Pitman-Yor prior for high order Markov models.
- Can now take $N \to \infty$, making use of *coagulation* and *fragmentation* properties of Pitman-Yor processes for computational tractability.
- Non-Markov model called the *sequence memoizer*.

[Goldwater et al. 2006a, Teh 2006b, Wood et al. 2009, Gasthaus et al. 2010]

Language Modelling

- Compare hierarchical Pitman-Yor model against hierarchical Dirichlet model, and two state-of-the-art language models (interpolated Kneser-Ney, modified Kneser-Ney).
- Results reported as perplexity scores.

T	N	IKN	MKN	HPYLM	HDLM
2e6	3	148.8	144.1	144.3	191.2
4e6	3	137.1	132.7	132.7	172.7
6e6	3	130.6	126.7	126.4	162.3
8e6	3	125.9	122.3	121.9	154.7
10e6	3	122.0	118.6	118.2	148.7
12e6	3	119.0	115.8	115.4	144.0
14e6	3	116.7	113.6	113.2	140.5
14e6	2	169.9	169.2	169.3	180.6
14e6	4	106.1	102.4	101.9	136.6

[Teh 2006b]

Compression

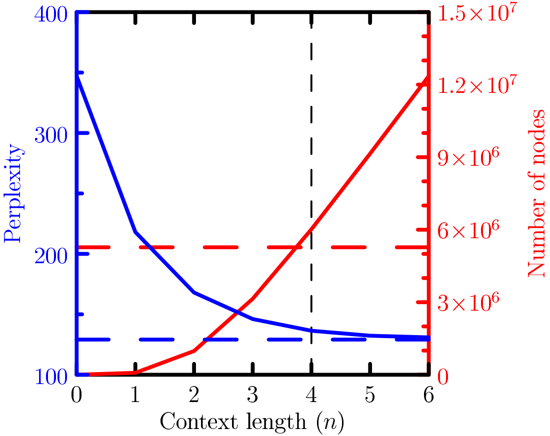
- Predictive models can be used to compress sequence data using entropic coding techniques.
- Compression results on Calgary corpus:

Model	Average bits / byte
`gzip`	2.61
`bzip2`	2.11
`CTW`	1.99
`PPM`	1.93
`Sequence Memoizer`	1.89

- See `http://deplump.com`.

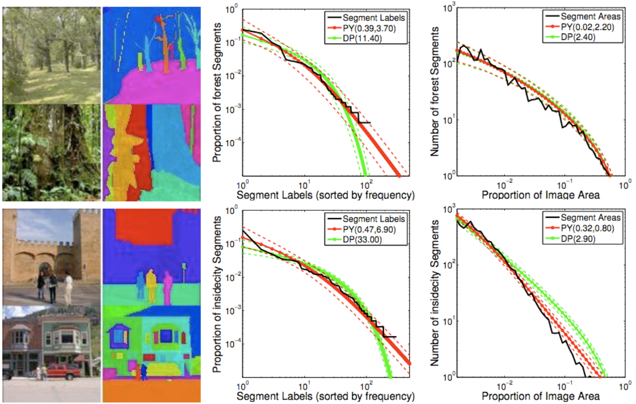
[Gasthaus et al. 2010]

Comparing Finite and Infinite Order Markov Models



[Wood et al. 2009]

Image Segmentation with Pitman-Yor Processes



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

[Sudderth and Jordan 2009]

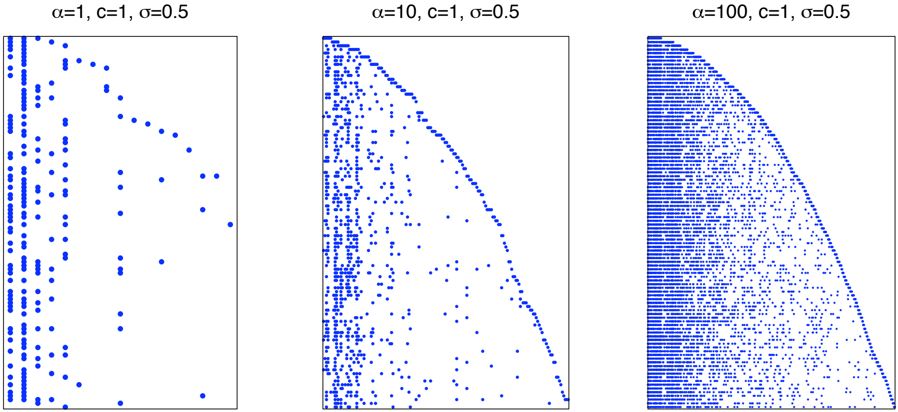
Stable Beta Process

- Extensions allow for different aspects of the generative process to be modelled:
  - α: controls the expected number of dishes picked by each customer.
  - c: controls the overall number of dishes picked by all customers.
  - σ: controls power-law scaling (ratio of popular dishes to unpopular ones).
- A completely random measure, with Lévy measure:

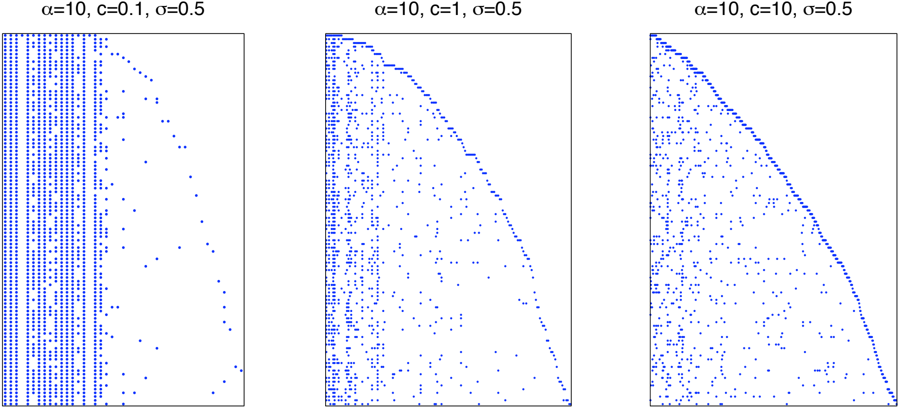
$$\alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \mu^{-\sigma-1}(1-\mu)^{c+\sigma-1} d\mu H(d\theta)$$

[Ghahramani et al. 2007, Teh and Görür 2009]

Stable Beta Process

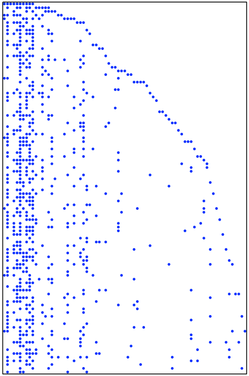


Stable Beta Process

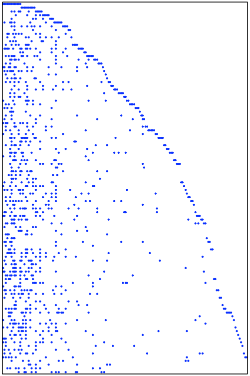


Stable Beta Process

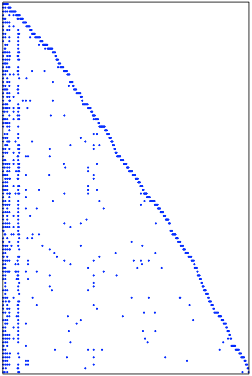
α=10, c=1, σ=0.2



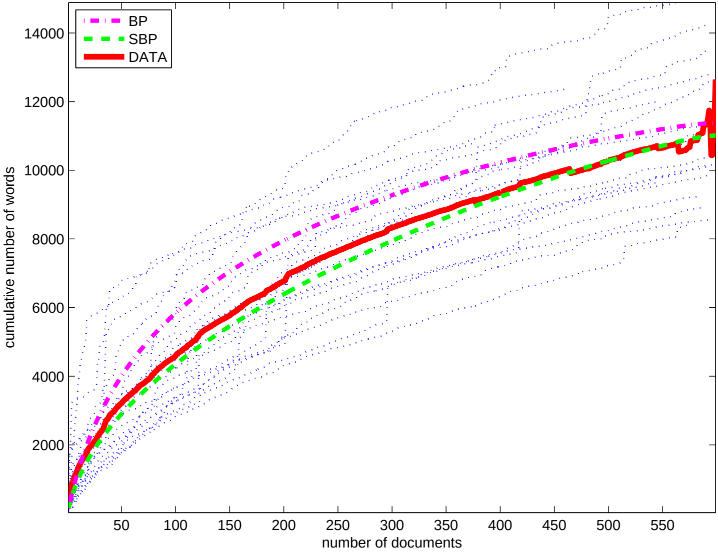
α=10, c=1, σ=0.5



α=10, c=1, σ=0.8



Modelling Word Occurrences in Documents



Outline

Summary

- Motivated Bayesian nonparametric modelling framework from a variety of applications.
- Sketched some of the more important theoretical concepts in building and working with such models.
- Missing from this tutorial: inference and computational issues, and asymptotic consistency and convergence.

[Hjort et al. 2010]

Thank You and Acknowledgements

- Cedric Archambeau
- Charles Blundell
- Hal Daume III
- Lloyd Elliott
- Jan Gasthaus
- Zoubin Ghahramani
- Dilan Görür
- Katherine Heller
- Lancelot James
- Michael I. Jordan
- Vinayak Rao
- Daniel Roy
- Jurgen Van Gael
- Max Welling
- Frank Wood

References I

- Adams, R. P., Wallach, H. M., and Ghahramani, Z. (2010). Learning the structure of deep sparse graphical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.
- Bart, E., Porteous, I., Perona, P., and Welling, M. (2008). Unsupervised learning of visual taxonomies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14.
- Bertoin, J. (2006). *Random Fragmentation and Coagulation Processes*. Cambridge University Press.
- 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., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the Association for Computing Machines*, 57(2):1–30.
- Chu, W., Ghahramani, Z., Krause, R., and Wild, D. L. (2006). Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *BIOCOMPUTING: Proceedings of the Pacific Symposium*.
- Doucet, A., de Freitas, N., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. New York: Springer-Verlag.
- 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*.
- Fox, E., Sudderth, E., Jordan, M. I., and Willsky, A. (2008). An HDP-HMM for systems with state persistence. In *Proceedings of the International Conference on Machine Learning*.

References II

- Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2010). Sharing features among dynamical systems with beta processes. In *Neural Information Processing Systems 22*. MIT Press.
- Gasthaus, J., Wood, F., and Teh, Y. W. (2010). Lossless compression based on the sequence memoizer. In *Data Compression Conference*.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian data analysis*. Chapman & Hall, London.
- 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., 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., 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.
- 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.
- Heller, K. A., Teh, Y. W., and Görür, D. (2009). Infinite hierarchical hidden Markov models. In *JMLR Workshop and Conference Proceedings: AISTATS 2009*, volume 5, pages 224–231.
- Hjort, N., Holmes, C., Müller, P., and Walker, S., editors (2010). *Bayesian Nonparametrics*. Number 28 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

References III

- 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.
- 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.
- Jordan, M. I. (2010). Hierarchical models, nested models and completely random measures. In *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. New York: Springer.
- 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.
- Liang, P., Petrov, S., Jordan, M. I., and Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Lijoi, A. and Pruenster, I. (2010). Models beyond the Dirichlet process. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics*. Cambridge University Press.
- MacKay, D. and Peto, L. (1994). A hierarchical Dirichlet language model. *Natural Language Engineering*.
- 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.
- Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

References IV

- 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. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.
- 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 Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.
- Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.
- 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. (2008). Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77.
- Teh, Y. W. (2006a). A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore.

References V

- 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, pages 1473–1480.
- Teh, Y. W. and Görür, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*, volume 22, pages 1838–1846.
- 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. (2010). Hierarchical Bayesian nonparametric models with applications. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics*. 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.
- 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, pages 564–571.
- Van Gael, J., Teh, Y. W., and Ghahramani, Z. (2009). The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 21, pages 1697–1704.
- Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.
- Wood, F., Archambeau, C., Gasthaus, J., James, L. F., and Teh, Y. W. (2009). A stochastic memoizer for sequence data. In *Proceedings of the International Conference on Machine Learning*, volume 26, pages 1129–1136.

References VI

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