RWTH AACHEN UNIVERSITY

Advanced Machine Learning
Lecture 17

Beta Processes II

07.01.2013

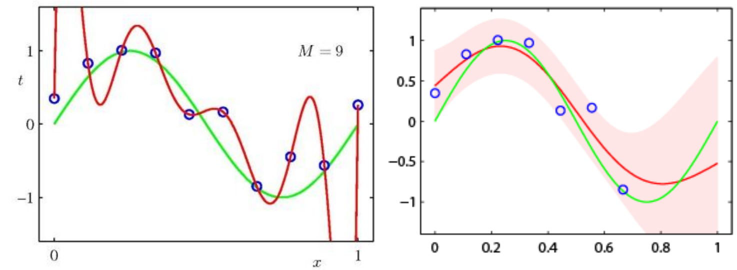
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

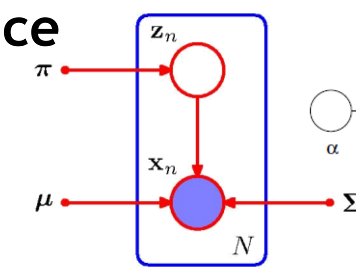
This Lecture: *Advanced Machine Learning*

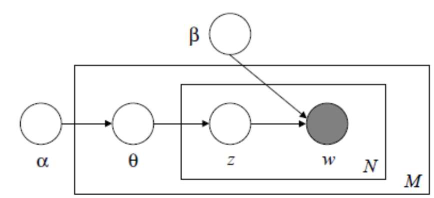
- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes

$$f : \mathcal{X} \rightarrow \mathbb{R}$$



- Bayesian Estimation & Bayesian Non-Parametrics
  - Prob. Distributions, Approx. Inference
  - Mixture Models & EM
  - Dirichlet Processes
  - Latent Factor Models
  - Beta Processes





- SVMs and Structured Output Learning
  - SV Regression, SVDD
  - Large-margin Learning

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

Topics of This Lecture

- Recap: Towards Infinite Latent Factor Models
  - General formulation
  - Finite latent feature model
  - Left-ordered binary matrices
  - Indian Buffet Process
- Beta Processes
  - Properties
  - Stick-Breaking construction
  - Inference
  - BPs for latent feature models
- Application: Nonparametric Hidden Markov Models
  - Graphical Model view
  - HDP-HMM
  - BP-HMM

Recap: Latent Factor Models

- **Mixture Models**
  - Assume that each observation was generated by *exactly* one of K components.
  - The uncertainty is just about which component is responsible.

- **Latent Factor Models**
  - Each observation is influenced by *each* of K components (factors or features) in a different way.
  - Sparse factor models: only a small subset of factors is active for each observation.

Recap: General Latent Factor Models

- General formulation
  - Assume that the data are generated by a noisy weighted combination of latent factors

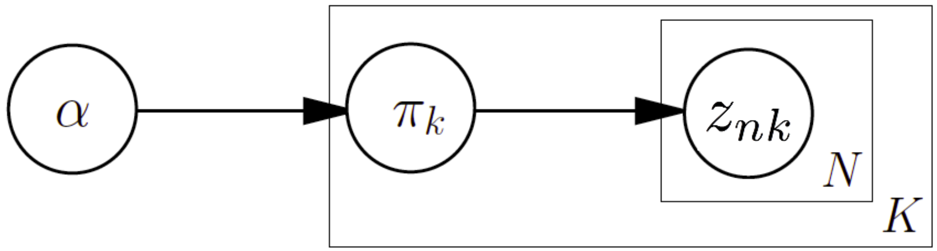
$$\mathbf{x}_n = \mathbf{F}\mathbf{y}_n + \epsilon$$

  - Mixture Models: DPs enforce that the main part of the probability mass is concentrated on few cluster components.
  - Latent Factor Models: enforce that each object is represented by *a sparse subset* of an unbounded number of features.

- Incorporating sparsity
  - Decompose $\mathbf{F}$ into the product of two components: $\mathbf{F} = \mathbf{Z} \otimes \mathbf{W}$, where $\otimes$ is the Hadamard product (element-wise product).
    - z_{mk} is a binary mask variable indicating whether factor k is "on".
    - w_{mk} is a continuous weight variable.

  $\Rightarrow$ Enforce sparsity by restricting the non-zero entries in $\mathbf{Z}$.

Recap: Finite Latent Feature Model



- **Probability model**
  - **Finite Beta-Bernoulli model**

$$\pi_k | \alpha \sim \text{Beta}(\frac{\alpha}{K}, 1)$$
$$z_{nk} | \pi_k \sim \text{Bernoulli}(\pi_k)$$

  - **Each z_{nk} is independent of all other assignments conditioned on π_k and the π_k are generated independently.**

Image source: Erik Sudderth

Towards Infinite Latent Feature Models

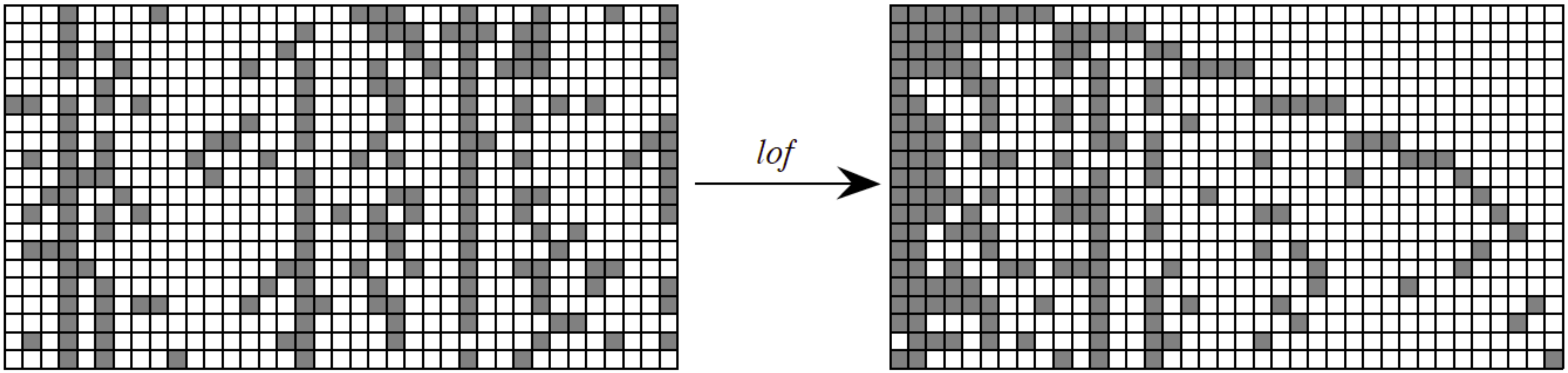
- Our goal is to let $K \to \infty$. Is this feasible with this model?
- Effective number of entries
  - We have shown: The expectation of the number of non-zero entries of $\mathbf{Z}$ is bounded by $N\alpha$, independent of K.

  $\Rightarrow$ $\mathbf{Z}$ is extremely sparse, only a finite number of factors is active.
- Probability for any particular matrix $\mathbf{Z}$
  - We have derived

$$p(\mathbf{Z}|\alpha) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}$$

  $\Rightarrow$ As $K \to \infty$, the probability of any particular $\mathbf{Z}$ will go to zero.
- Solution: Define equivalence classes of matrices

Recap: Equivalence Class of Binary Matrices



- **Equivalence class of binary matrices**
  - Define a function $\mathrm{lof}(\mathbf{Z})$ that maps binary matrices into left-ordered binary matrices by ordering the columns of $\mathbf{Z}$ by the magnitude of the binary number expressed by that column.
  - There is a unique left-ordered form for every binary matrix.
  - Two matrices $\mathbf{Y}$ and $\mathbf{Z}$ are equivalent iff $\mathrm{lof}(\mathbf{Y}) = \mathrm{lof}(\mathbf{Z})$.
  - The lof-equivalence class of $\mathbf{Z}$ is denoted $[\mathbf{Z}]$.

Image source: Zoubin Ghahramani

Towards Infinite Latent Feature Models

- **Taking the limit $K \to \infty$**
  - **Probability of a** lof**-equivalence class of binary matrices**

$$p([\mathbf{Z}]|\alpha) = \sum_{\mathbf{Z}\in[\mathbf{Z}]} p(\mathbf{Z}|\alpha) = \frac{K!}{\prod_{h=0}^{2^N-1} K_h!} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}$$

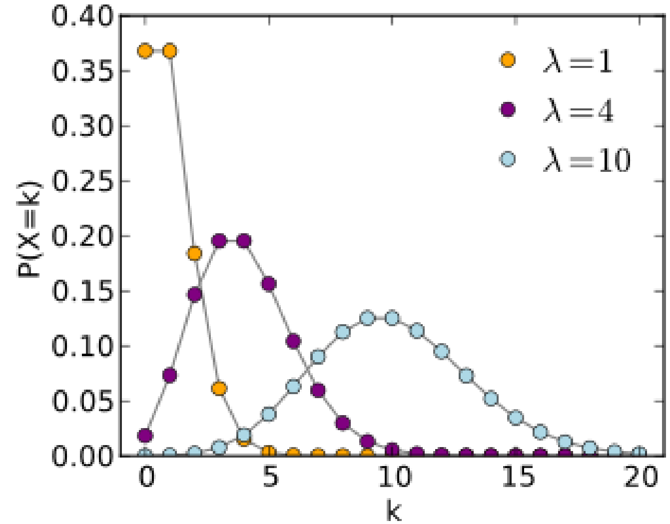
  - **Reordering the columns such that $m_k > 0$ if $k \le K_+$, and $m_k = 0$ otherwise, we can derive (after several intermediate steps)**

$$\lim_{K\to\infty} p([\mathbf{Z}]|\alpha) \;=\; \frac{\alpha^{K_+}}{\prod_{h=0}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}$$

  - **where H_N is the N^{th} harmonic number $H_N = \sum_{j=1}^{N} 1/j$.**
  - **Again, this distribution is exchangeable.**

Excursion: The Poisson Distribution

- **Motivation**
  - *Express the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate λ and independently of the time since the last event.*



- **Definition**
  - **Probability mass function for discrete Variable X**

$$p(X = k) = \mathrm{Pois}(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, ...$$

  - **Properties**

$$\mathbb{E}[x] = \mathrm{Var}[x] = \lambda$$

  - **The Poisson distribution can be derived as the limit of a Binomial distribution.**

Image source: Wikipedia

Excursion: The Poisson Distribution

- Derivation (Law of rare events)
  - Consider an interval (e.g., in time or space) in which events happen at random with known average number λ.
  - Divide the interval in N subintervals $I_1,\ldots,I_N$ of equal size.

  $\Rightarrow$ The probability that an event will fall into subinterval I_k is λ/N.
  - Consider the occurrence of an event in I_k to be a Bernoulli trial.
  - The total number of events X will then be Binomial distributed with parameters N and λ/N.

$$p(X=k) = \mathrm{Bin}(k;N,\lambda/N) = \frac{N!}{k!(N-k)!}\left(\frac{\lambda}{N}\right)^k \left(1-\frac{\lambda}{N}\right)^{N-k}$$

  - For large N, this can be approximated by a Poisson distribution

$$\lim_{N\to\infty} p(X=k) = \lim_{N\to\infty} \frac{N(N-1)\ldots(N-k+1)}{N^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{N}\right)^N\left(1-\frac{\lambda}{N}\right)^{-k}$$

$$= 1 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot 1$$

Why Poisson?

- Why are we interested in Poisson distributions?
  1. We have Bernoulli trials for the individual z_{nk} and are interested in the infinite limit the resulting model.

  2. Compare the result we just derived for the infinite latent feature model

$$\lim_{K \to \infty} p([\mathbf{Z}]|\alpha) = \frac{\alpha^{K_+}}{\prod_{h=0}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

with the definition of a Poisson distribution

$$\text{Pois}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

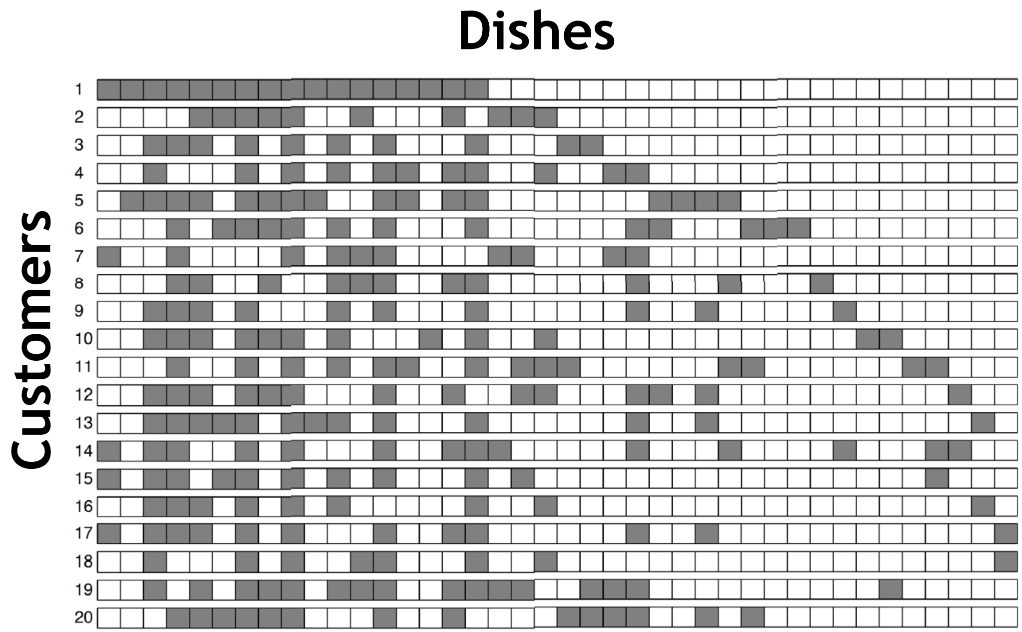
$\Rightarrow$ There is clearly some Poisson distributed component, but the exact connection is hard to grasp due to the complex notation.

- *We will see the connection more clearly in the following...*

Topics of This Lecture

- Recap: Towards Infinite Latent Factor Models
  - General formulation
  - Finite latent feature model
  - Left-ordered binary matrices
  - Indian Buffet Process
- Beta Processes
  - Properties
  - Stick-Breaking construction
  - Inference
  - BPs for latent feature models
- Application: Nonparametric Hidden Markov Models
  - Graphical Model view
  - HDP-HMM
  - BP-HMM

The Indian Buffet Process

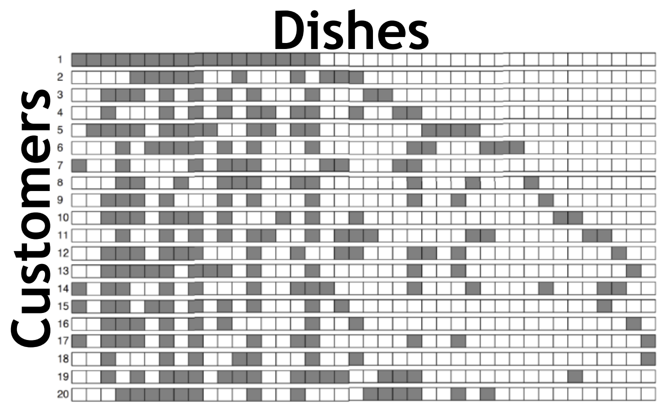


"Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes"

[Zoubin Ghahramani]

Image source: Zoubin Ghahramani

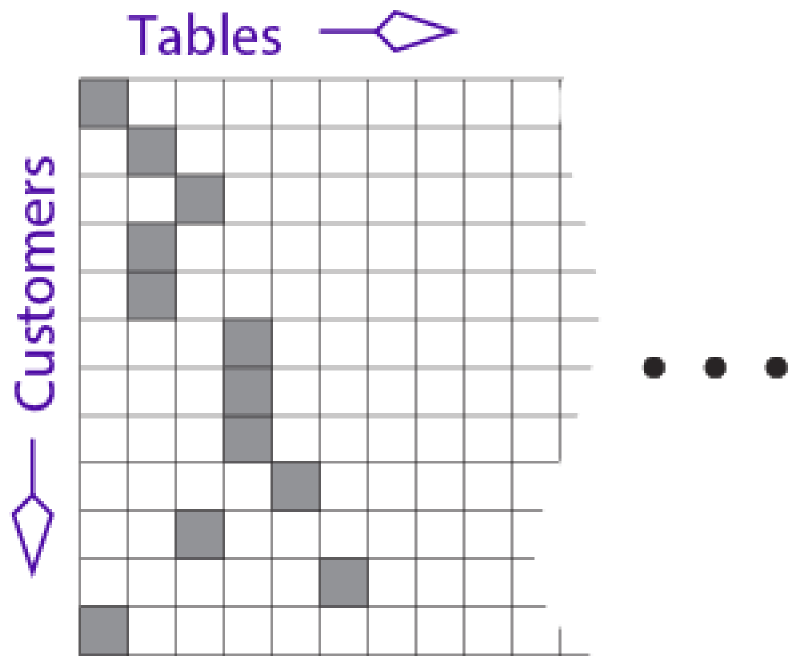
The Indian Buffet Process

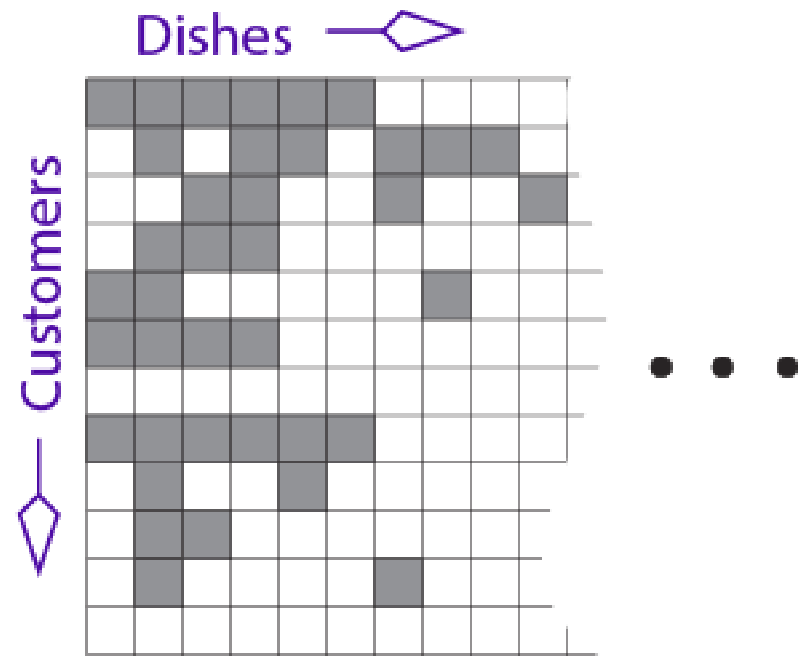


- **Analogy to Chinese Restaurant Process**
  - Visualize feature assignment as a sequential process of customers sampling dishes from an (infinitely long) buffet.
  - 1st customer starts at the left of the buffet, and takes a serving from each dish, stopping after a $\text{Poisson}(\alpha)$ number of dishes as her plate becomes overburdened.
  - The n^{th} customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself with probability m_k/n, and trying a $\text{Poisson}(\alpha/n)$ number of new dishes.
  - The customer-dish matrix is our feature matrix, $\mathbf{Z}$.

Image source: Zoubin Ghahramani

Comparison: CRP vs. IBP





Chinese Restaurant Process

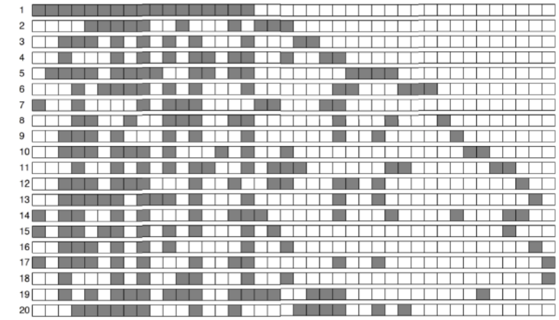
- Each customer is assigned to a single component.
- *Tables* correspond to **mixture components**.

Indian Buffet Process

- Each customer can be assigned to multiple components.
- *Dishes* correspond to **latent factors/features**.

Image source: Yee Whye Teh

The Indian Buffet Process (IBP)



- Analysis
  - Let $K_1^{(n)}$ indicate the number of new dishes sampled by customer n. It can be shown that the probability of any particular matrix $\mathbf{Z}$ being produced is

$$p(\mathbf{Z}|\alpha) = \frac{\alpha^{K_+}}{\prod_{n=1}^{N} K_1^{(n)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

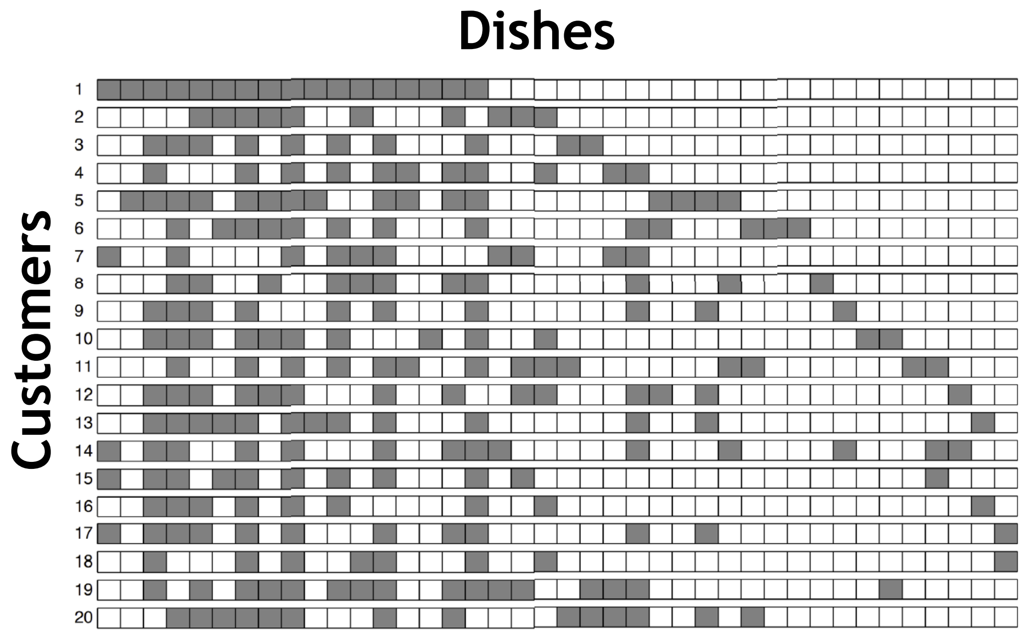
  - The matrices generated by the IBP are generally not in lof, but they are also not ordered arbitrarily, since new dishes are always added to the right.
  - If we only pay attention to the lof-equivalence class $[\mathbf{Z}]$, we obtain the exchangeable distribution

$$p([\mathbf{Z}]|\alpha) = \frac{\alpha^{K_+}}{\prod_{h=0}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}$$

$\Rightarrow$ ***Same result as for the infinite latent feature model!***

Image source: Zoubin Ghahramani

The Indian Buffet Process (IBP)



- **Properties of the IBP**
  - Generative process to create samples from an infinite latent feature model.
  - The IBP is infinitely exchangeable, up to a permutation of the order with which dishes are listed in the feature matrix.
  - The number of features sampled at least once is $\mathcal{O}(\alpha \log N)$.

Image source: Zoubin Ghahramani

The Indian Buffet Process (IBP)

- More properties
  1. ***The effective dimension of the model,*** K_+***, follows a*** $\mathrm{Poisson}(\alpha H_N)$ ***distribution.***

     Proof: Easily shown, since $K_+ = \sum_n \mathrm{Poisson}(\alpha/n)$**.**

  2. ***The number of features possessed by each object follows a*** $\mathrm{Poisson}(\alpha)$ ***distribution.***

     Proof: The 1st customer chooses a $\mathrm{Poisson}(\alpha)$ **number of dishes. By exchangeability, this also holds for all other customers.**

  3. ***The expected number of non-zero entries in*** $\mathbf{Z}$ ***is*** $N\alpha$***.***

     Proof: This directly follows from the previous result.

  4. ***The number of non-zero entries in*** $\mathbf{Z}$ ***will follow a*** $\mathrm{Poisson}(N\alpha)$ ***distribution.***

     Proof: Follows from properties of sums of Poisson variables.

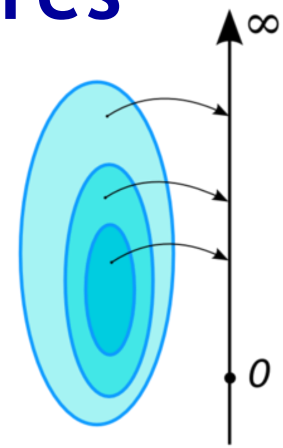
Topics of This Lecture

- Recap: Towards Infinite Latent Factor Models
  - General formulation
  - Finite latent feature model
  - Left-ordered binary matrices
  - Indian Buffet Process
- **Beta Processes**
  - **Properties**
  - **Stick-Breaking construction**
  - **Inference**
  - **BPs for latent feature models**
- Application: Nonparametric Hidden Markov Models
  - Graphical Model view
  - HDP-HMM
  - BP-HMM

The Beta Process

- IBP and Exchangeability
  - Since the IBP is infinitely exchangeable, De Finetti's theorem states that it must have an underlying random measure.
  - The Beta Process is the De Finetti random measure for the IBP, just like the DP was the De Finetti random measure for the CRP.
- Beta Processes
  - Just like the DP, the Beta Process is a distribution on distributions.
  - A formal definition would require an excursion into the theory of completely random measures, which is mostly beyond the scope of this lecture.
  - *In the following, I will therefore only highlight its most important properties...*

Excursion: Completely Random Measures



- **Measure**
  - A **measure** on a set is a systematic way to assign a number to each suitable subset of that set.
- **Completely random** means
  - The random variables obtained by evaluating the random measure on disjoint subsets of the probability space are **mutually independent**.
  - Draws from a completely random measure are **discrete** (up to a fixed deterministic component).

  $\Rightarrow$ Thus, we can represent such a draw as a weighted collection of atoms on some probability space, as we did for the DP.
- **Sidenote**
  - The DP is not a completely random measure, since its weights are constrained to sum to one. Thus, the independence assumption does not hold for the DP!

Image source: Wikipedia

Beta Process

- **Formal definition**
  - A **Beta Process** $B \sim \mathrm{BP}(c, \alpha H)$ is a **completely random** discrete measure of the 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

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

  - The Beta Process with $c = 1$ is the **De Finetti measure for the IBP**. (For $c \neq 1$, we get a 2-parameter generalization of the IBP).

Beta Process

- **Less formal definition**
  - **Define the random measure B as a set of weighted atoms $\{\theta_k{}^*\}$**

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

where $\mu_k \in (0,1)$ and the atoms $\{\theta_k{}^*\}$ are drawn from a base measure H_0 on Θ.

  - **We define the Beta Process as a *distribution on distributions* (analogously to the DP) for random measures with weights between 0 and 1 and denote it by $B \sim \mathrm{BP}(\alpha,\, H_0)$.**

- **Notes**
  - **The weights μ_k *do not* sum to 1 $\Rightarrow$ B is not a probability measure**
  - **A Beta Process *does not* have Beta distributed marginals!**

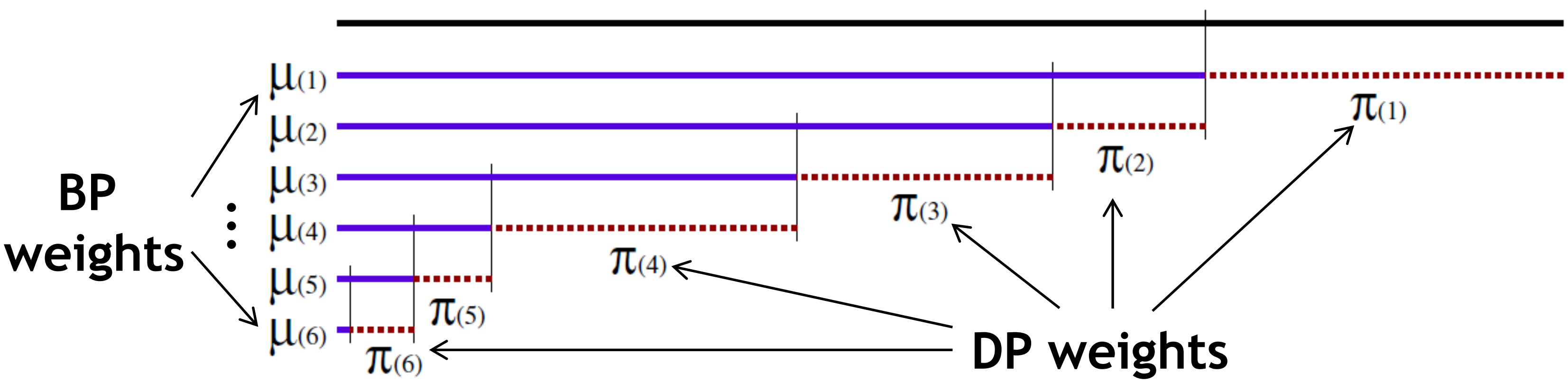
Stick-Breaking Construction for BPs

- Explicit construction of the BP
  - For $c = 1$, there is a closed-form Stick-Breaking Process

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

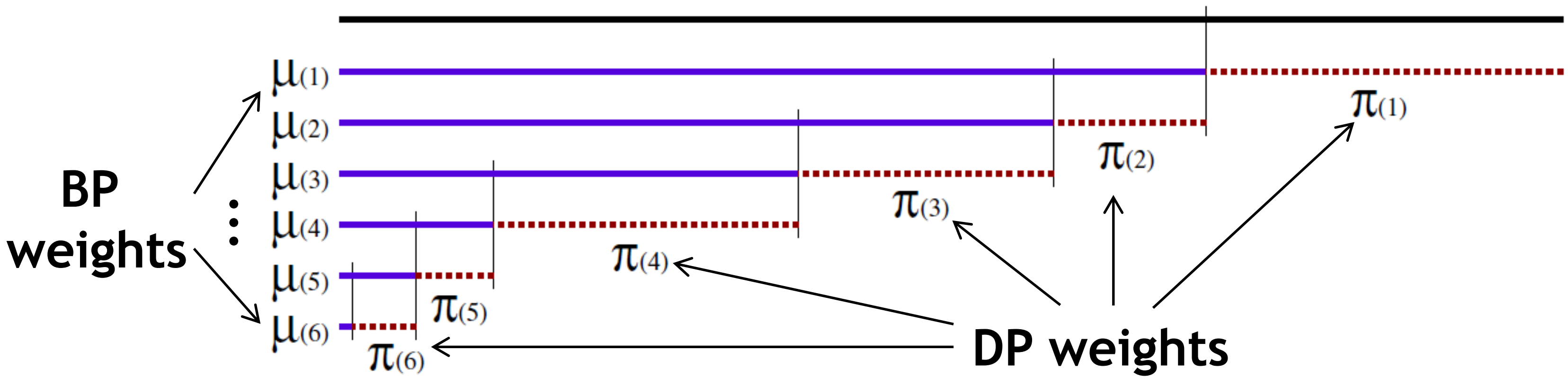
$$\mu_k = (1 - \beta_k) \prod_{l=1}^{k-1} (1 - \beta_l) \qquad \theta_k^* \sim H_0 \qquad B = \sum_{k=1}^{\infty} \mu_k \delta_{\theta_k^*}$$

  - This is the complement of the Stick-Breaking Process for DPs!



  - As for the DP, we write this procedure as $\mu_k, \theta_k^* \sim \text{GEM}(\alpha, H_0)$

Stick-Breaking Construction for BPs



- **Interpretation**
  - **The DP weights can be thought off as portions broken off an initially unit-length stick.**
  - **The BP weights then correspond to the remaining stick length.**
- **Properties**
  - **DP: stick lengths sum to one and are not monotonically decreasing (only on average).**
  - **BP: stick lengths do not sum to one and are decreasing.**

Inference for Beta Processes

- Goal
  - Infer the posterior distribution of the latent features

$$p(\mathbf{Z}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})$$

  - As for the DP, exact inference is intractable, since the normalization requires a sum over all possible binary matrices $\mathbf{Z}$.

- Approximate Inference
  - Inference in BPs can be performed using either the IBP or the Stick-Breaking construction.
  - A number of algorithms have been proposed using MCMC or variational approximations. Since the BP is typically part of a larger model, many of those algorithms are however too complex to present here.
  - Given posterior samples of $\mathbf{Z}$, one typically examines the highest-probability sample (the MAP estimate) to get a sense of the latent feature structure.

Gibbs Sampling for the IBP

- Simple approach: Gibbs Sampling
  - In order to specify a Gibbs sampler, we need to derive the full conditional distribution

$$p(z_{nk} = 1 | \mathbf{Z}_{-(n,k)}, \mathbf{X}) \;\propto\; p(\mathbf{X}|\mathbf{Z}) p(z_{nk} = 1 | \mathbf{Z}_{-(n,k)})$$

  where $\mathbf{Z}_{-(n,k)}$ denotes the entries of $\mathbf{Z}$ other than z_{nk}.

  - The likelihood term $p(\mathbf{X}|\mathbf{Z})$ depends on the model chosen for the observed data.
  - The conditional assignments $p(z_{nk} \,|\, \mathbf{z}_{-n,k})$ can be derived from the exchangeable IBP. Choosing an ordering such that the n^{th} object corresponds to the last customer, we obtain

$$p(z_{nk} = 1 | \mathbf{z}_{-n,k}) \;=\; \frac{m_{-n,k}}{N}$$ for any k such that $m_{-n,k} > 0$.

  - Similarly, the number of new features associated with object n should be drawn from a $\mathrm{Poisson}(\alpha/N)$ distribution.

Source: [Ghahramani et al., 2006]

Topics of This Lecture

- Recap: Towards Infinite Latent Factor Models
  - General formulation
  - Finite latent feature model
  - Left-ordered binary matrices
  - Indian Buffet Process
- Beta Processes
  - Properties
  - Stick-Breaking construction
  - Inference
  - BPs for latent feature models
- Application: Nonparametric Hidden Markov Models
  - Graphical Model view
  - HDP-HMM
  - BP-HMM

BPs and Latent Feature Models

- **Building a Latent Feature Model from the BP**
  - **Define a new random measure**

$$X_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\theta_k^*}$$

where $z_{nk} \sim \mathrm{Bernoulli}(\mu_k)$**.**

  - **The random measure** X_n **is then said to be distributed according to a Bernoulli Process with the Beta Process as its base measure:** $X_n \sim \mathrm{BeP}(B), \quad B \sim \mathrm{BP}(\alpha, H_0)$**.**
  - **A draw from the Bernoulli Process places unit mass on those atoms for which** $z_{nk} = 1$**; this defines, which latent features are "on" for the** n^{th} **observation.**
  - N **draws from the Bernoulli Process yield an IBP-distributed binary matrix Z [Thibaux & Jordan, 2007].**

Source: [Gershman & Blei, 2012; Paisley & Carin, 2009]]

Application: BP Factor Analysis

- Recap: Factor Analysis
  - Goal: Model a data matrix, $\mathbf{X} \in \mathbb{R}^{D \times N}$, as the multiplication of two matrices, $\mathbf{\Phi} \in \mathbb{R}^{D \times K}$ and $(\mathbf{W} \otimes \mathbf{Z}) \in \mathbb{R}^{K \times N}$, plus an error matrix $\mathbf{E}$.

$$\mathbf{X} = \mathbf{\Phi}(\mathbf{W} \otimes \mathbf{Z}) + \mathbf{E}$$

  - Or written in vector notation for each observation $\mathbf{x}_n$

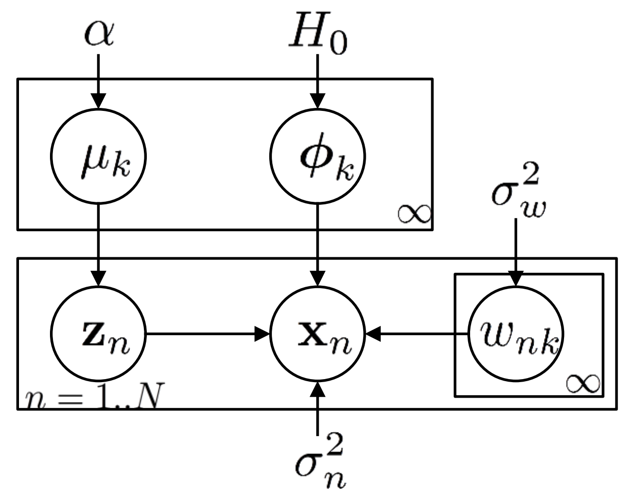
$$\mathbf{x}_n = \mathbf{\Phi}(\mathbf{w}_n \otimes \mathbf{z}_n) + \boldsymbol{\epsilon}_n$$

- Basic idea of BP-FA
  - Model the matrices $\mathbf{\Phi}$ and $\mathbf{Z}$ as N draws from a Bernoulli Process, parameterized by a Beta Process $B \sim \mathrm{BP}(\alpha, H_0)$ with a multivariate Normal distribution as its base measure H_0.

Source: [Gershman & Blei, 2012; Paisley & Carin, 2009, 2010]

Application: BP Factor Analysis

- **Graphical Model**



- **Possible BP-FA realization**
  - Draw the weight vector $\mathbf{w}_n$ from a Gaussian prior.
  - Draw the atoms ϕ_k and their weights μ_k from the Beta Process (e.g., using the stick-breaking construction).
  - Construct each $\mathbf{z}_n$ by turning on a subset of these atoms according to a draw from the Bernoulli Process.
  - Generate the noisy observation $\mathbf{x}_n$

$$\mathbf{w}_n \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$$

$$\mu_k \sim \mathrm{GEM}(\alpha)$$

$$\phi_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

$$z_{nk} \sim \mathrm{Bernoulli}(\mu_k)$$

$$\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$$

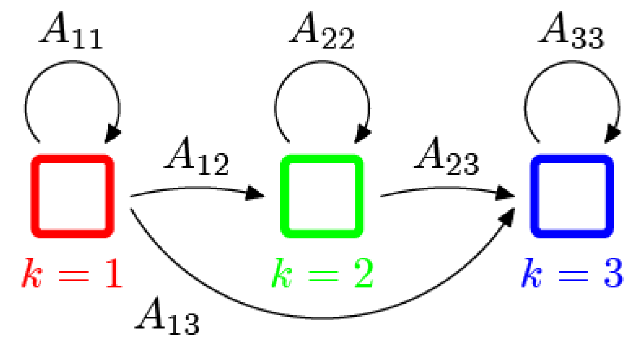
$$\mathbf{x}_n = \mathbf{\Phi}(\mathbf{w}_n \otimes \mathbf{z}_n) + \boldsymbol{\epsilon}_n$$

Topics of This Lecture

- Recap: Towards Infinite Latent Factor Models
  - General formulation
  - Finite latent feature model
  - Left-ordered binary matrices
  - Indian Buffet Process
- Beta Processes
  - Properties
  - Stick-Breaking construction
  - Inference
  - BPs for latent feature models
- Application: Nonparametric Hidden Markov Models
  - Graphical Model view
  - HDP-HMM
  - BP-HMM

Hidden Markov Models (HMMs)

- Probabilistic model for sequential data
  - Widely used in speech recognition, natural language modeling, handwriting recognition, financial forecasting,…
- Traditional view:
  - Finite state machine
  - Elements:
    - State transition matrix $\mathbf{A}$,
    - Production probabilities $p(\mathbf{x} \mid k)$.



- Graphical model view
  - Dynamic latent variable model
  - Elements:
    - Observation at time n: $\mathbf{x}_n$
    - Hidden state at time n: $\mathbf{z}_n$
    - Conditionals $p(\mathbf{z}_{n+1}|\mathbf{z}_n)$, $p(\mathbf{x}_n|\mathbf{z}_n)$

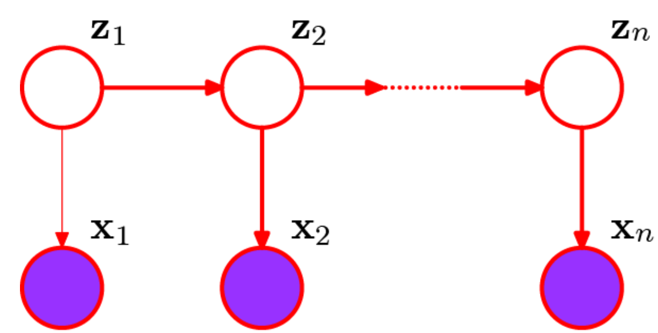
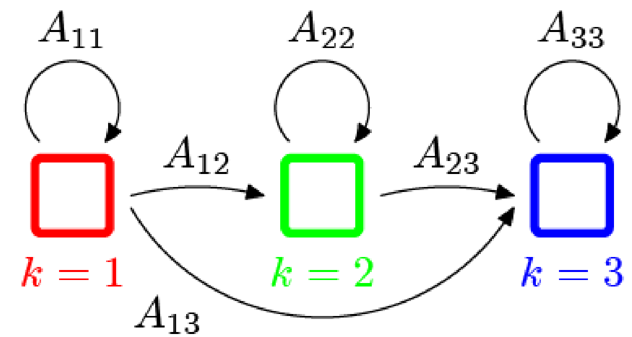


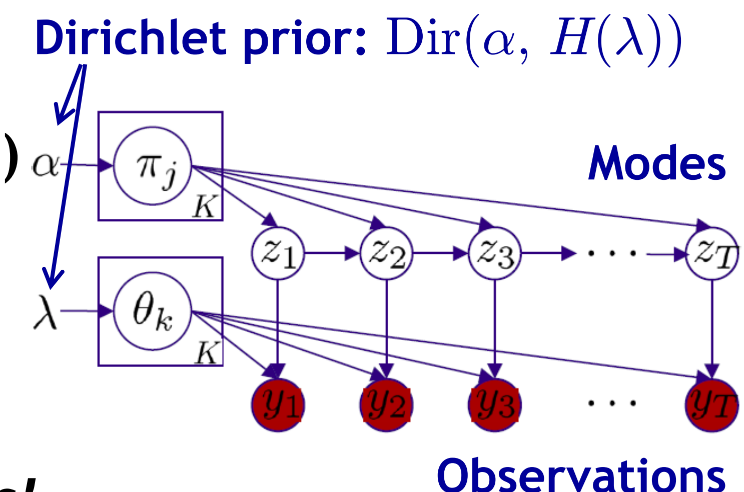
Image source: C.M. Bishop, 2006

Hidden Markov Models (HMMs)

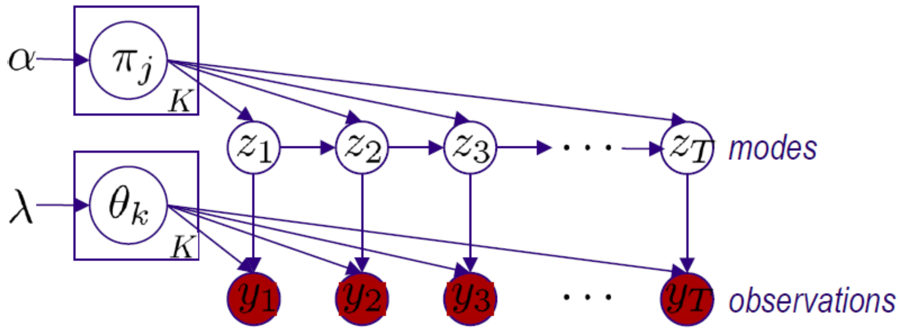
- Traditional HMM learning
  - Each state has a distribution over observable outputs $p(\mathbf{x} \mid k)$, e.g., modeled as a Gaussian.
  - Learn the output distributions together with the transition probabilities using an EM algorithm.



- Graphical Model view
  - Treat the HMM as a mixture model
  - Each state is a component ("mode") in the mixture distribution.
  - From time step to time step, the responsible component switches according to the transition model.
  - *Advantage: we can introduce priors!*

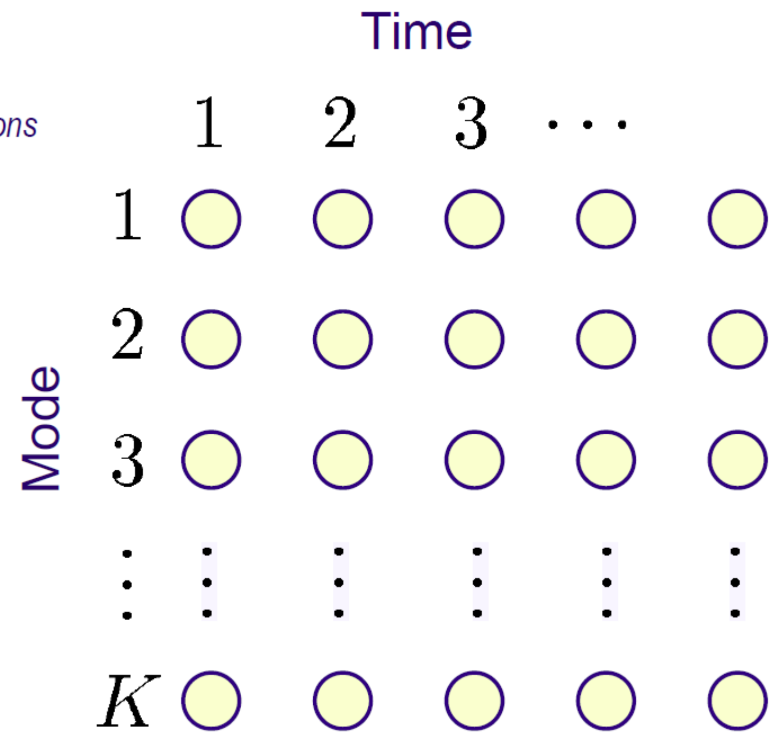


HMM: Mixture Model View

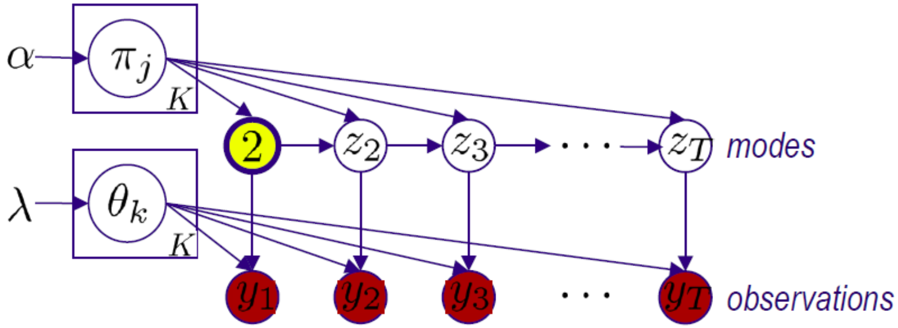


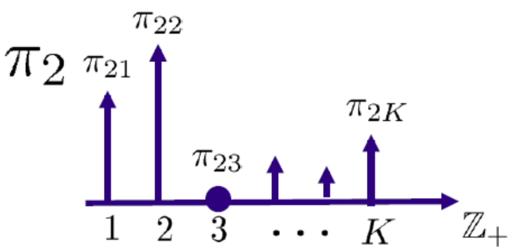
$$z_t \sim \pi_{z_{t-1}}$$
$$y_t \sim F(\theta_{z_t})$$

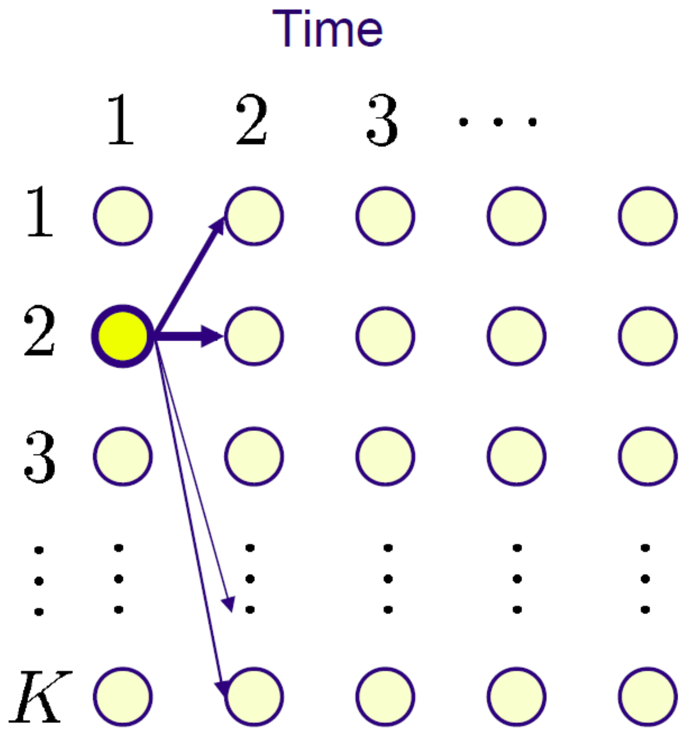
$$P = \begin{bmatrix} \text{———} & \pi_1 & \text{———} \\ \text{———} & \pi_2 & \text{———} \\ & \vdots & \\ \text{———} & \pi_K & \text{———} \end{bmatrix}$$



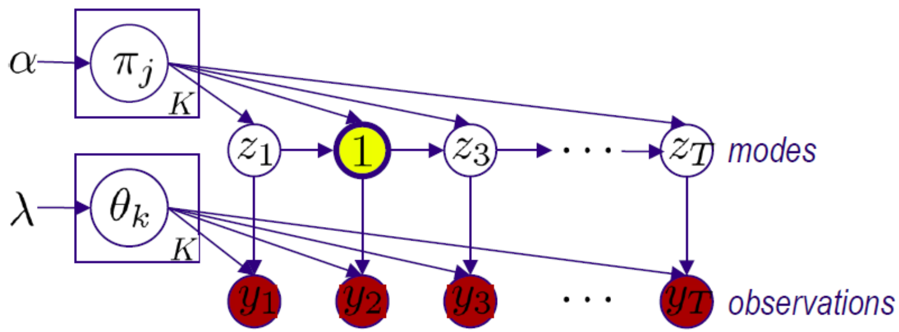
HMM: Mixture Model View

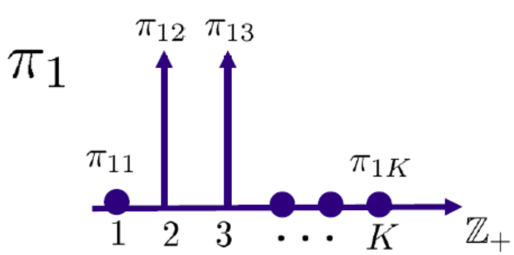


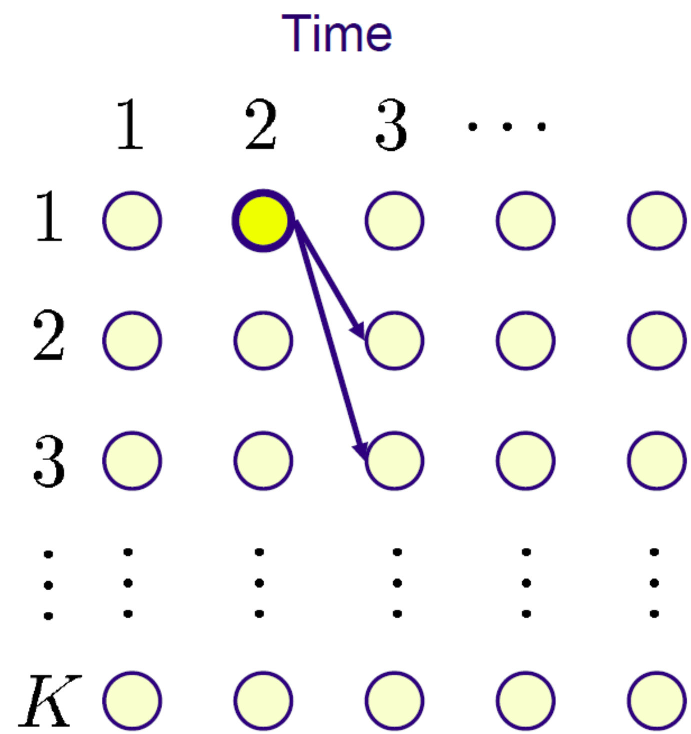




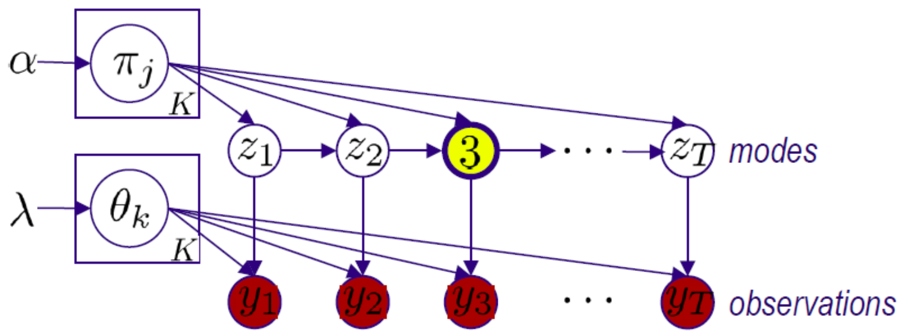
HMM: Mixture Model View

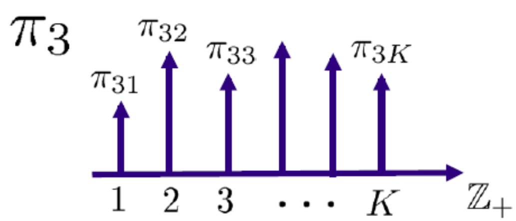


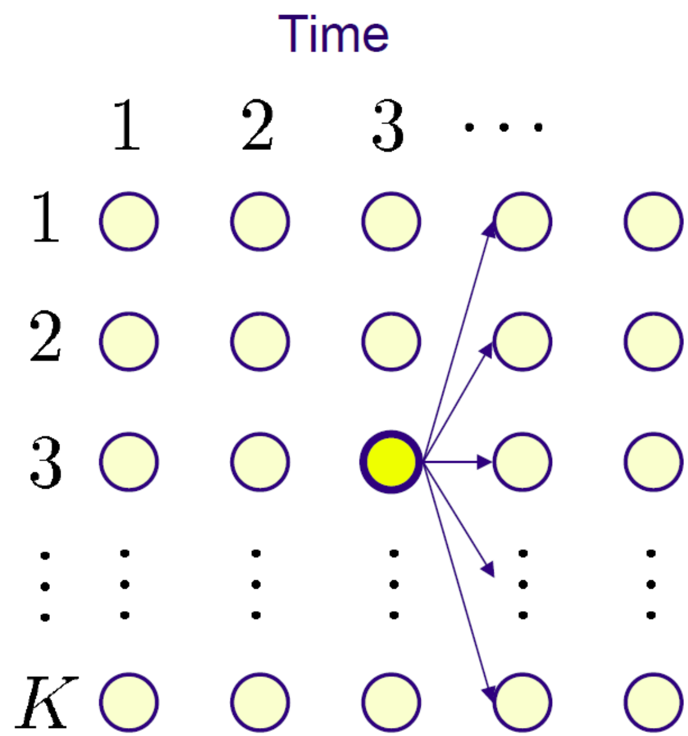




HMM: Mixture Model View



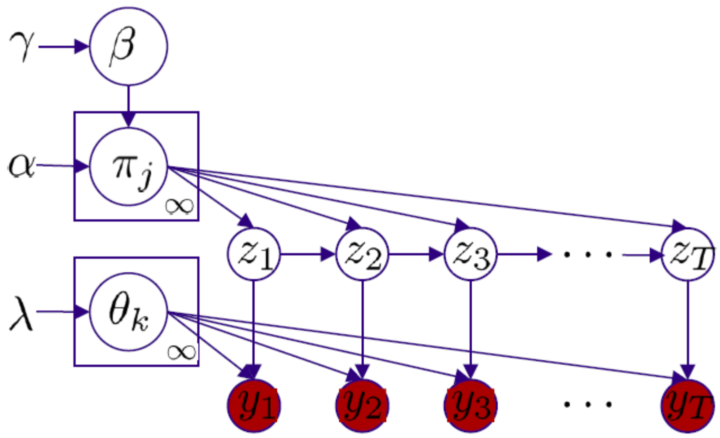




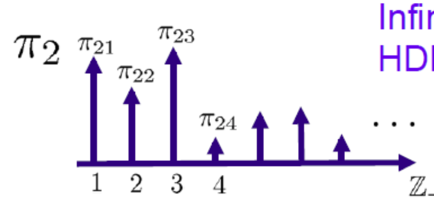
Important issue: How many modes?

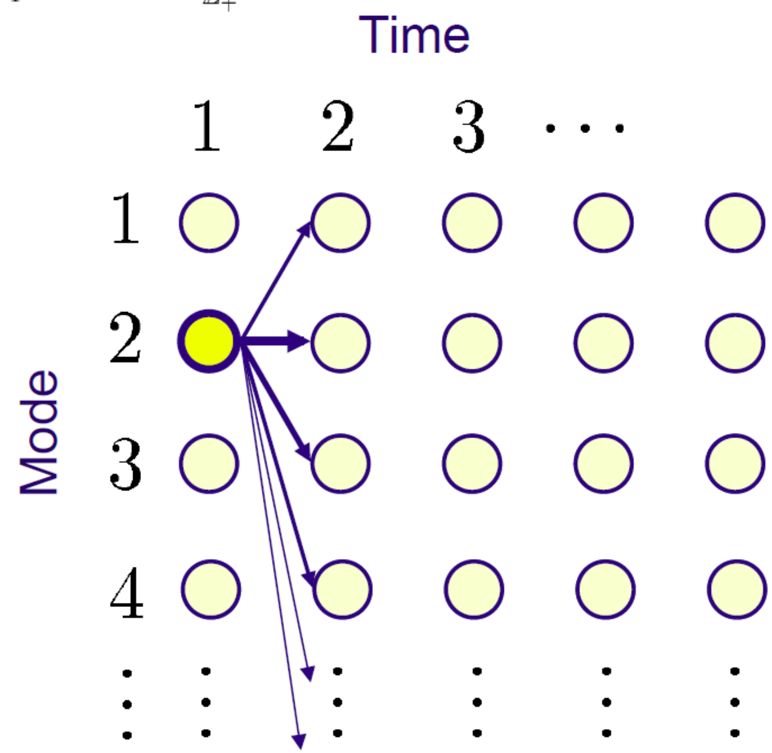
Hierarchical Dirichlet Process HMM

Infinite HMM: Beal, et.al., *NIPS* 2002
HDP-HMM: Teh, et. al., *JASA* 2006



HDP HMM





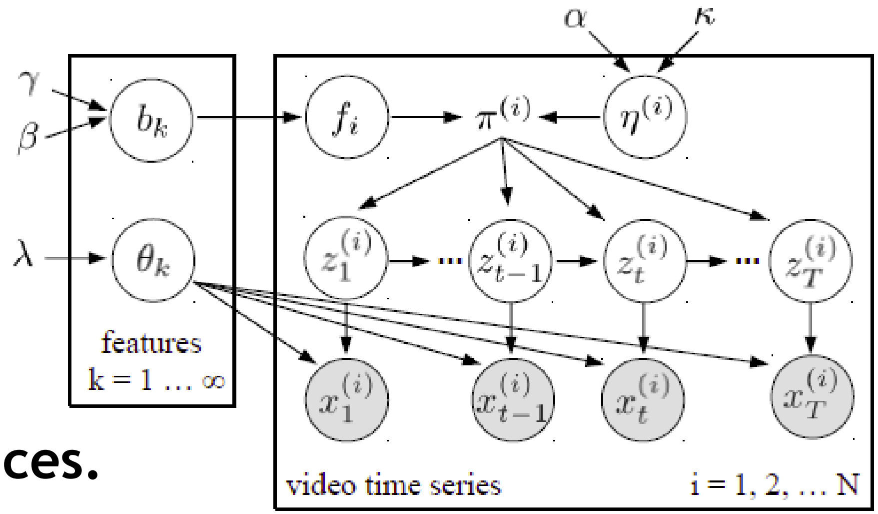
Infinite state space

- **Dirichlet Process**
  - Mode space of unbounded size
  - Model complexity adapts to observations
- **Hierarchical DP**
  - Ties mode transition distributions
  - *Shared* sparsity

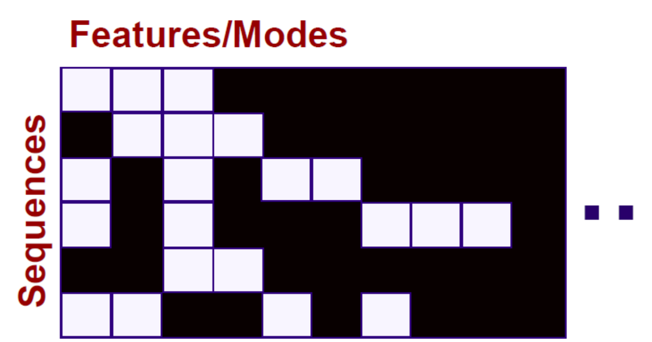
RWTH AACHEN UNIVERSITY

Beta Process HMM

- Goal: Transfer knowledge between related time series
  - E.g., activity recognition in video collections
  - Allow each system to switch between an arbitrarily large set of dynamical modes ("behaviors").
  - Share behaviors across sequences.



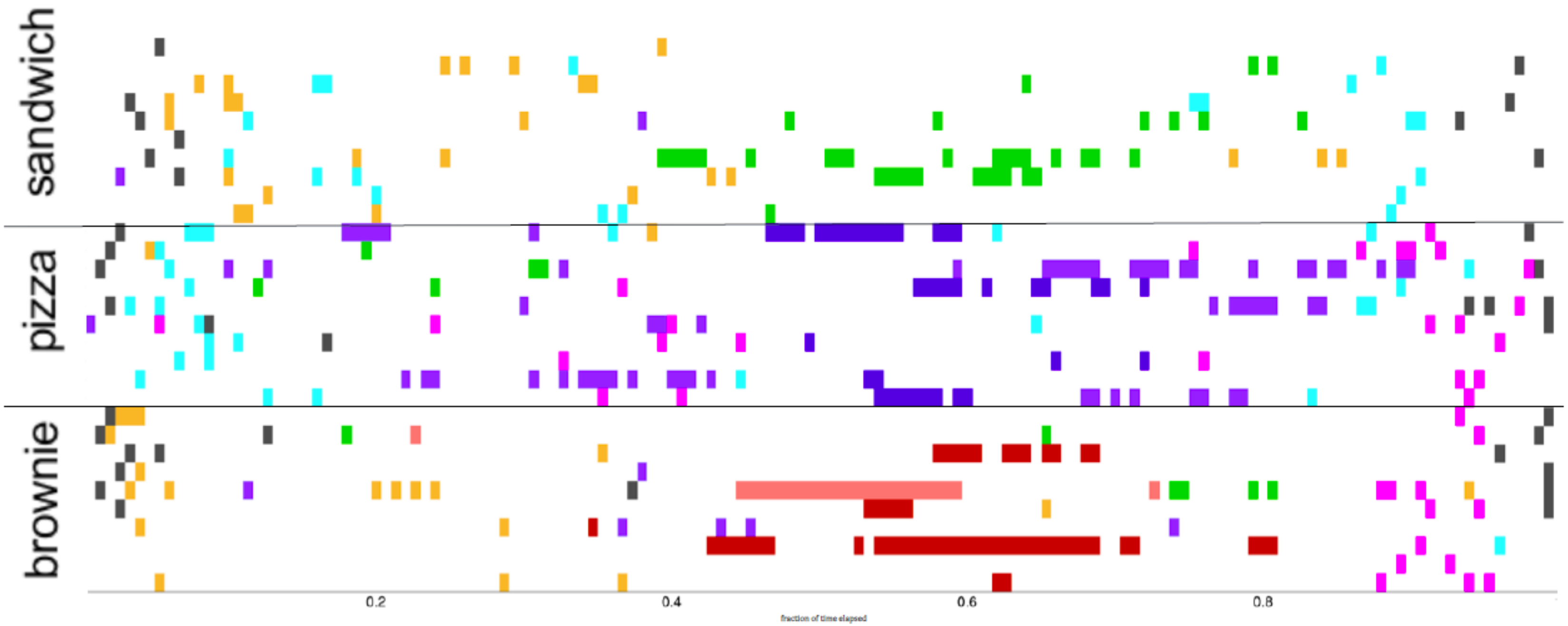
- Beta Processes enforce sparsity
  - HDPs would force all videos to have non-zero probability of displaying all behaviors.
  - Beta Processes allow a video to contain only a sparse subset of relevant behaviors.



[Hughes & Sudderth, 2012]

Image source: Erik Sudderth

Unsupervised Discovery of Activity Patterns



CMU Kitchen dataset

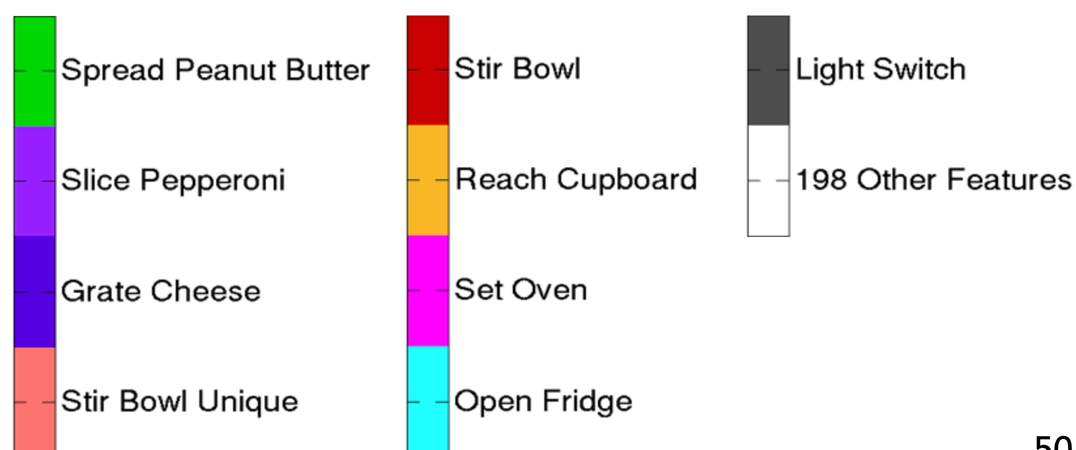


Image source: Erik Sudderth

Summary

- Beta Processes
  - Powerful nonparametric framework for latent feature models
  - Much younger than the DP, so much is still in development.
  - E.g., stick-breaking construction was only shown in 2010.
  - Beta Processes and the IBP can be used in concert with different likelihood models in a variety of applications.

- Many other applications being developed, e.g.
  - Infinite Independent Component Analysis
  - Matrix factorization for collaborative filtering (recommender systems)
  - Latent causal discovery for medical diagnosis
  - Protein complex discovery
  - …

References and Further Reading

- Tutorial papers for infinite latent factor models
  - A good introduction to the topic
    - Z. Ghahramani, T.L. Griffiths, P. Sollich, "Bayesian Nonparametric Latent Feature Models", Bayesian Statistics, 2006.
  - A tutorial on Hierarchical BNPs, including Beta Processes
    - Y.W. Teh, M.I. Jordan, Hierarchical Bayesian Nonparametric Models with Applications. Bayesian Nonparametrics, Cambridge Univ. Press, 2010.
- Example applications of BPs
  - BP Factor Analysis
    - J. Paisley, F. Carin, Nonparametric Factor Analysis with Beta Process Priors, ICML 2009.
  - BP-HMMs for discovery of activity patterns
    - M.C. Hughes, E.B. Sudderth, Nonparametric Discovery of Activity Patterns from Video Collections. CVPR Workshop on Perceptual Organization in Computer Vision, 2012.