Bayesian nonparametrics

Approches bayésiennes non paramétriques

François Caron

Department of Statistics, Oxford

Ecole d'été de Peyresq en traitement du signal et des images
July 2016

Outline

Outline

Data models

- Model-based statistical methods
- Definition of a statistical model describing the data generating process
- Based on an *interpretation of the data*, motivated by the problem at hand, and not an *explanation of the data*.

  "Essentially, all models are wrong, but some are useful."

  George E.P. Box

- Necessary reduction of the problem, oriented to the problem to solve

Data models

- Bayesian methods
- Probability distribution of the data $m(y)$

$$m(y) = \int_{\Phi} \pi(\phi, y) d\phi$$

  where $\phi \in \Phi$ denotes the set of parameters of the model, which are themselves treated as random variables.
- Bayesian data modeling: specification of $\pi(\phi, y)$
- Graphical models

Inference

- Posterior distribution

$$\pi(\phi|y) = \frac{\pi(\phi, y)}{m(y)}$$

  which represents the uncertainty on the model parameters given the data.
- Various numerical methods
  - Markov Chain Monte Carlo
  - Sequential Monte Carlo
  - Variational Bayes methods

Building Bayesian data model

- Construction of $\boldsymbol{\pi(\phi, y)}$ dictated by several antagonistic desiderata
  - Fit to the data
  - Predictive power
  - Elegance and simplicity; existence of remarkable statistical properties
  - Interpretability of the parameters
  - Simplicity and automaticity of inference
  - Computational tractability and scalability
- Key point: model complexity, related to the number of parameters
  - Too simple model will suffer from under-fitting and have poor predictive performances
  - Too complicated model will loose in interpretability and computational tractability

Bayesian nonparametrics

- Bayesian parametrics: $\boldsymbol{dim(\phi) < \infty}$
- Bayesian nonparametrics: $\boldsymbol{dim(\phi) = \infty}$
- Advantages
  - Distribution of the data has a wider support than that provided by a parametric model
  - Model complexity increases with the number of data
  - Robust and adaptive framework
  - Conjugacy: Inference algorithms often as simple as for parametric models
  - Interesting statistical properties: power-law behavior, sparsity
- Limitations
  - Requires more advanced mathematical tools (stochastic processes)
  - Some counter-examples for consistency of Bayesian estimators with BNP priors

Bayesian nonparametrics

Historical background

- Stochastic processes used in a Bayesian framework: Dirichlet processes (Ferguson, 1973), Gaussian processes (O'Hagan 1978), beta processes (Hjort, 1990), Polya tree priors (Lavine, 1990) but applications rather limited
- With the development of MCMC algorithms in the early 90's, those models can now be used in hierarchical models
  - MCMC for Dirichlet process mixture models (Escobar and West, 1995)
- Increased interest in statistics and machine learning, with the development of novel models, algorithms and applications
- Now standard tools of the Bayesian toolbox
  - A workshop every two years in statistics
  - A workshop on average every two years in machine learning

Bayesian nonparametrics

Rough cartography of BNP models

Application	Basic model	More advanced/flexible models
Clustering Density estimation	Dirichlet Process	Pitman-Yor, normalized CRMs, Poisson-Kingman, Polya trees, log-Gaussian processes dependent DP, hierarchical DP, Nested DP
Latent feature	Beta process	Stable BP, dependent BP, GGP-Poisson
Hidden Markov models	HMM-HDP	'sticky' HDP-HMM, reversible HMM
Regression	Gaussian process	DPMs and others
Survival analysis	Beta processes	Neutral to the right processes

Outline

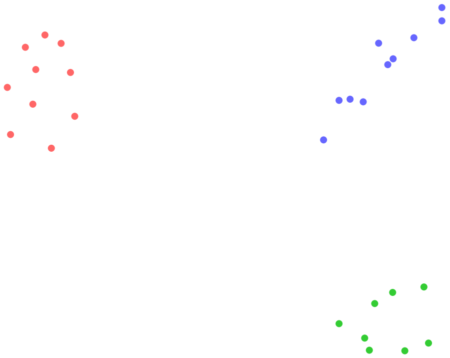
Introduction
Clustering

- Cluster/partition a set of items $i = 1, \dots, n$ into clusters

Introduction

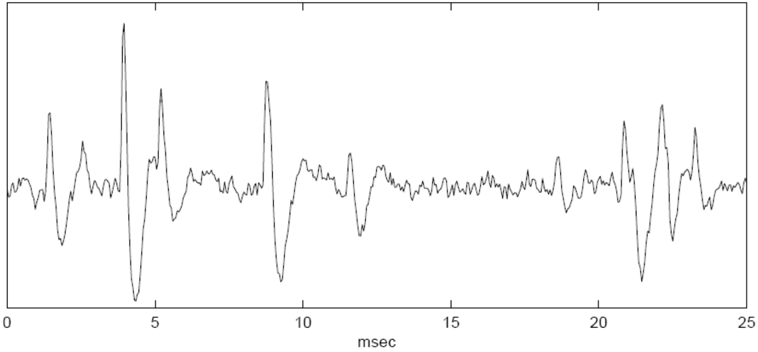
Clustering

- Cluster/partition a set of items $i = 1, \dots, n$ into clusters



Example: Spike sorting

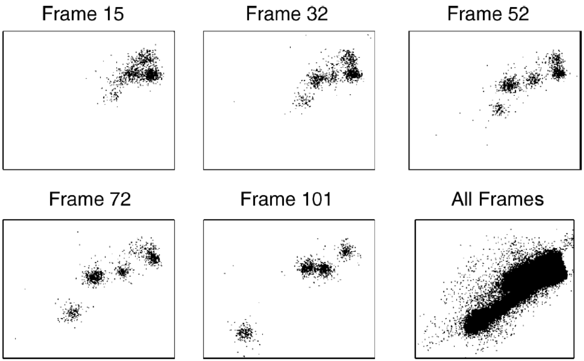
- Brief voltage spikes recorded by a microelectrode
- Goal: Sort signals to assign particular spikes to putative neurons
- Unknown number of neurons, background noise



[Bar-Hillel et al., 2006, Gasthaus et al., 2008]

Example: Spike sorting

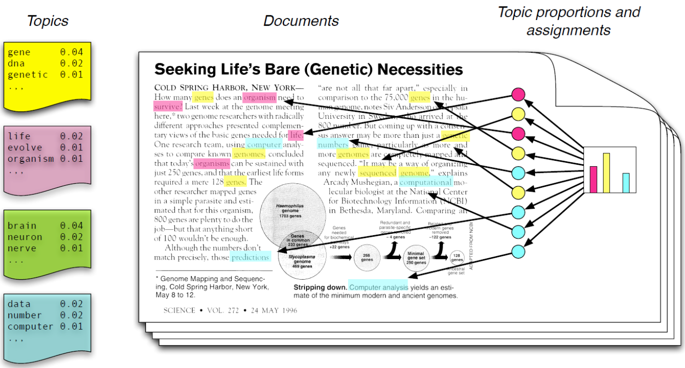
- Brief voltage spikes recorded by a microelectrode
- Goal: Sort signals to assign particular spikes to putative neurons
- Unknown number of neurons, background noise



[Bar-Hillel et al., 2006, Gasthaus et al., 2008]

Example: Topic modeling

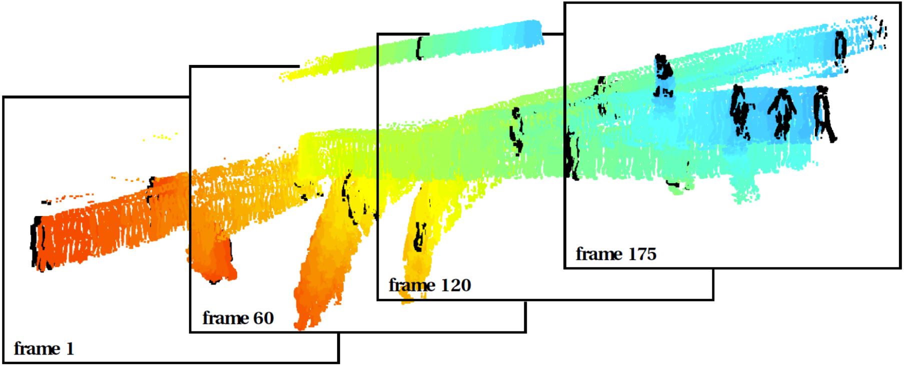
- Words in documents
- Objective: find topics within documents
- 'Bag of words' assumption within documents



[Blei, 2012]

Example: Multiple-object tracking

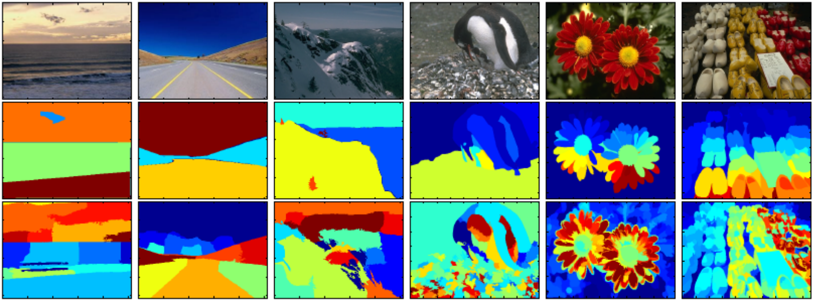
- Track an unknown and varying number of objects over time
- Joint data association and tracking problem



[Caron et al., 2016]

Example: Image segmentation

- Segment an image into homogeneous regions



[Xu et al., 2016, Sodjo et al., 2016]

Introduction

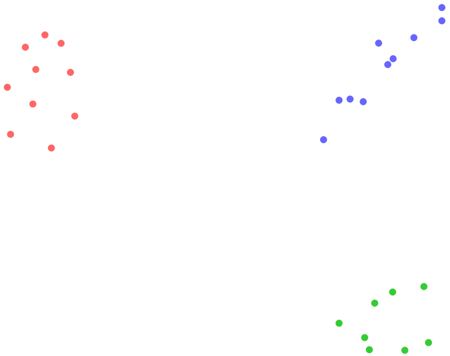
Clustering

- Cluster/partition a set of items $i = 1, \ldots, n$ into clusters

Introduction

Clustering

- Cluster/partition a set of items $i = 1, \ldots, n$ into clusters



Introduction

Clustering

- Partition
$$\Pi_n = \{A_{n,1}, \ldots, A_{n,K_n}\}$$
where $A_{n,j}$, $j = 1, \ldots, K_n$ non-empty and non-overlapping subsets of $[n] := \{1, \ldots, n\}$ with $\cup_{j=1}^{K_n} A_{n,j} = [n]$
- $A_{n,j}$ are clusters, $K_n \leq n$ is the number of clusters
- Example
$$\Pi_6 = \{\{1,4,5\},\{2,3\},\{6\}\}$$

- Notations: often convenient to represent the partition using allocation variables, e.g.
$$(c_1 = 1, c_2 = 2, c_3 = 2, c_4 = 1, c_5 = 1, c_6 = 3)$$

⚠ The cluster labels are irrelevant!
$$(c_1 = 3, c_2 = 1, c_3 = 1, c_4 = 3, c_5 = 3, c_6 = 2)$$
encode the same partition

- Model-based: f_U defines the parametric shape of a cluster
  - Example: f_U is a Gaussian where $U = (\mu, \Sigma)$ is the mean and covariance matrix of that Gaussian
- Cluster locations U_j, $j = 1, \ldots, K_n$
- Partition Π_n of the data
- Likelihood

$$p(y_1, \ldots, y_n | U_{1:K_n}, \Pi_n) = \prod_{j=1}^{K_n} \prod_{i \in A_j} f_{U_j}(y_i)$$

- Bayesian approach: (U_j) and Π_n treated as random variables
- Nonparametric approach: K_n can increase unboundedly with the number of items n
- Exchangeable random partition
  - For any n, the distribution is invariant w.r.t. any permutation of $[n]$, e.g.

$$\Pr(\{\{1,2\},\{3\}\}) = \Pr(\{\{2,3\},\{1\}\}) = \Pr(\{\{1,3\},\{2\}\})$$

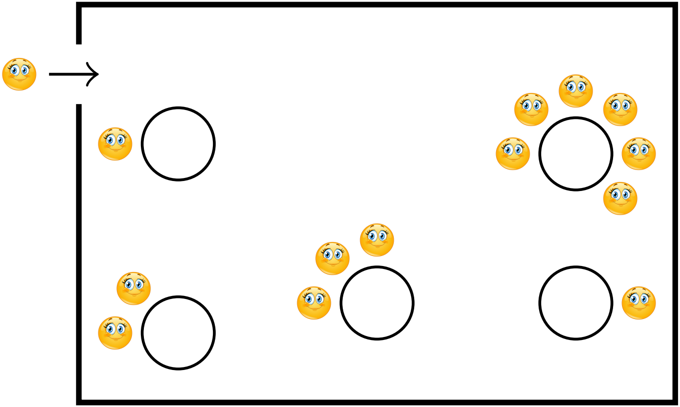
  - Labelling/ordering of the items is of no importance

- Assume additionally that

$$\Pr(c_{n+1} = \text{new}|c_1, \ldots, c_n) = f(n) \tag{1}$$

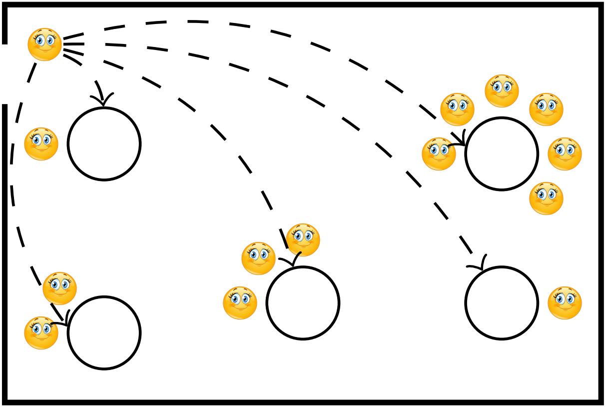
  i.e. the probability of creating a new cluster only depends on the sample size n (and not on the cluster sizes nor the number of clusters)
- The two properties of exchangeability and (1) characterize a class of partition models
- Chinese restaurant process: generative process for this class of exchangeable partitions

Chinese restaurant process



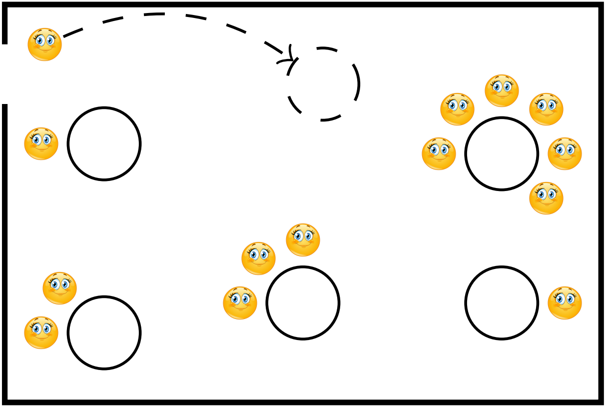
- Customer $n+1$
  - Joins an existing table $j = 1, \ldots, K_n$ w.p. $\frac{m_{n,j}}{n+\alpha}$
  - Sits at a new table w.p. $\frac{\alpha}{n+\alpha}$

Chinese restaurant process



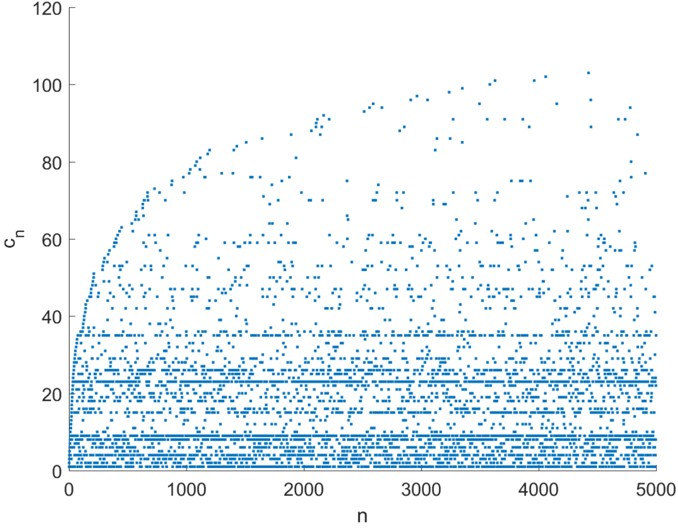
- Customer $n+1$
  - Joins an existing table $j = 1, \ldots, K_n$ w.p. $\frac{m_{n,j}}{n+\alpha}$
  - Sits at a new table w.p. $\frac{\alpha}{n+\alpha}$

Chinese restaurant process

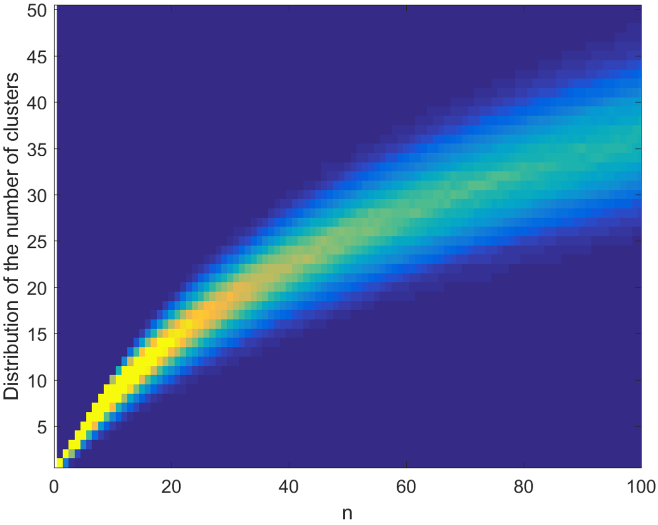


- Customer $n+1$
  - Joins an existing table $j = 1, \ldots, K_n$ w.p. $\frac{m_{n,j}}{n+\alpha}$
  - Sits at a new table w.p. $\frac{\alpha}{n+\alpha}$

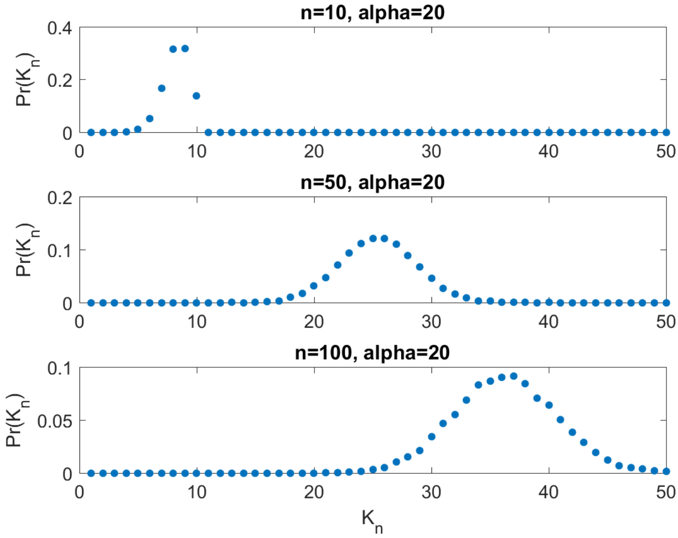
Chinese restaurant Process



Chinese restaurant Process



Chinese restaurant Process



Chinese restaurant process

- Rich-gets-richer process

$$\Pi_n \sim \text{CRP}(\alpha, n)$$

- Parameter $\alpha > 0$
- Logarithmic growth of the number of clusters

$$\mathbb{E}[K_n] = \sum_{i=0}^{n-1} \frac{\alpha}{\alpha + i}$$

$$\frac{K_n}{\alpha \log n} \to 1 \text{ almost surely as } n \to \infty$$

Hierarchical model

$$\Pi_n \sim \text{CRP}(\alpha, n)$$

for $j = 1, \ldots, K_n$,

$$U_j \sim G_0$$

For $i = 1, \ldots, n$

$$y_i | \Pi_n, U_1, \ldots, U_{K_n} \sim f_{U_{c_i}}$$

Posterior inference

- Conjugate DPM model

$$p(y_{1:n}|\Pi_n) = \prod_{j=1}^{K_n} q_{A_{n,j}}(y_{1:n})$$

  where

$$q_A(y_{1:n}) = \int_\Theta \prod_{i \in A} f_\theta(y_i) G_0(d\theta)$$

  can be computed analytically.
- Marginal posterior

$$\Pr(\Pi_n|y_{1:n})$$

- Gibbs sampler
  - At each iteration
    - For $i = 1, \ldots, n$, sample $c_i|c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n, y_{1:n}$

Posterior inference

- Let $\Pi_{-i} = \{A_{-i,1}, \ldots, A_{-i,K_{-i}}\}$ be the partition of $[n]\backslash\{i\}$ obtained by removing item i from Π_n, and $m_{-i,j}$ the size of the clusters $j = 1, \ldots, K_{-i}$
- By exchangeability, for $j = 1, \ldots, K_{-i}$,

$$\Pr(c_i = j|\Pi_{-i}) = \frac{m_{-i,j}}{\alpha + n - 1}$$

and

$$\Pr(c_i = \text{new}|\Pi_{-i}) = \frac{\alpha}{\alpha + n - 1}$$

- Full conditional

$$\Pr(c_i = j|\Pi_{-i}, y_{1:n}) \propto m_{-i,j} \frac{q_{A_{-i,j}\cup\{i\}}(y_{1:n})}{q_{A_{-i,j}}(y_{1:n})}$$

$$\Pr(c_i = \text{new}|\Pi_{-i}, y_{1:n}) \propto \alpha q_{\{i\}}(y_{1:n})$$

Dirichlet distribution

- Distribution on the $d-1$ simplex

$$(\pi_1, \ldots, \pi_d) \sim \text{Dirichlet}(a_1, \ldots, a_d)$$

where $\pi_j \geq 0$, $\sum_{j=1}^d \pi_j = 1$, $a_j > 0$.

- Density (w.r.t. to the Lebesgue measure on the $d-1$ simplex)

$$p(\pi_1, \pi_2, \ldots, \pi_{d-1}) = \frac{\Gamma(\sum_{j=1}^d a_j)}{\prod_{j=1}^d \Gamma(a_j)} \prod_{j=1}^d \pi_j^{a_j - 1}$$

where $\pi_j \geq 0$, $\sum_{j=1}^{d-1} \pi_j \leq 1$ and $\pi_d = 1 - \sum_{j=1}^{d-1} \pi_j$.

Dirichlet distribution

- Parametrization

$$a_j = \alpha p_{0j}$$

where $\alpha > 0$ and $\sum_{j=1}^{d} p_{0j} = 1$.

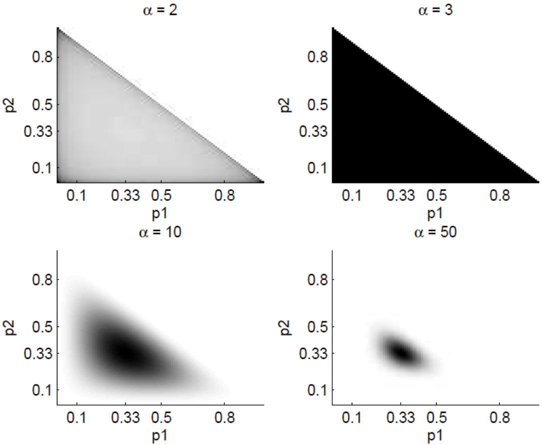
- Properties

$$\mathbb{E}[\pi_j] = p_{0j}$$

$$\mathbf{Var}[\pi_j] = \frac{p_{0j}(1 - p_{0j})}{1 + \alpha}$$

Dirichlet distribution

$d = 3$, $p_0 = (1/3, 1/3, 1/3)$



Dirichlet distribution

- Let $z_i \in \{1, \ldots, d\}$ be categorical random variables such that

$$\Pr(z_i = j | \pi_{1:d}) = \pi_j$$

- Let $m_{n,j} = \text{card}\{i = 1, \ldots, n | z_i = j\}$

$$\Pr(z_{1:n} | \pi_{1:d}) = \prod_{j=1}^{d} \pi_j^{m_{n,j}}$$

- Conjugacy

$$(\pi_1, \ldots, \pi_d) | z_{1:n} \sim \text{Dirichlet}(\underbrace{\alpha p_{01} + m_{n,1}}_{\widetilde{\alpha}\widetilde{p}_{01}}, \ldots, \underbrace{\alpha p_{0d} + m_{n,d}}_{\widetilde{\alpha}\widetilde{p}_{0d}})$$

where $\widetilde{\alpha} = \alpha + n$ and $\widetilde{p}_{0j} = \frac{m_{n,j}}{\alpha+n} + \frac{\alpha}{\alpha+n} p_{0j}$

Dirichlet distribution

- Predictive

$$\Pr(z_{n+1} = j | z_{1:n}) = \frac{\alpha p_{0j} + m_{n,j}}{\alpha + n}$$

- Proof

$$\begin{aligned}
\Pr(z_{n+1} = j | z_{1:n}) &= \mathbb{E}_{\pi_{1:d}|z_{1:n}}[\Pr(z_{n+1} = j | \pi_{1:d}, z_{1:n})] \\
&= \mathbb{E}_{\pi_{1:d}|z_{1:n}}[\Pr(z_{n+1} = j | \pi_{1:d})] \\
&= \mathbb{E}_{\pi_{1:d}|z_{1:n}}[\pi_j]
\end{aligned}$$

Dirichlet Process

- Distribution over probability distributions on $\boldsymbol{\Theta}$

$$G \sim \mathrm{DP}(\alpha, G_0)$$

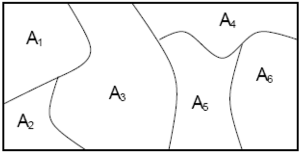
where
  - G_0 is the base probability distribution
  - $\alpha > 0$ is the scale parameter

Definition

For all partition $A_1, ..., A_d$ of $\boldsymbol{\Theta}$

$$(G(A_1), ..., G(A_d)) \sim \mathrm{Dirichlet}\left(\alpha G_0(A_1), ..., \alpha G_0(A_d)\right)$$

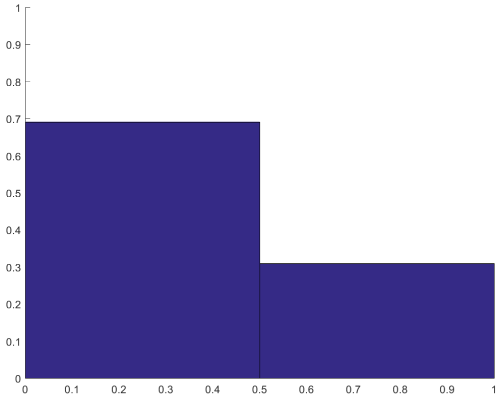
where $\mathrm{Dirichlet}(b_1, \ldots, b_d)$ is the standard Dirichlet distribution.



[Ferguson, 1973]

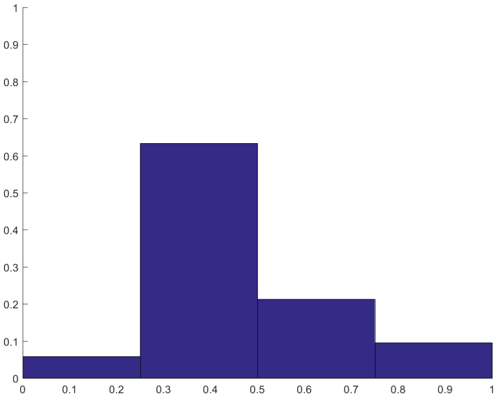
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



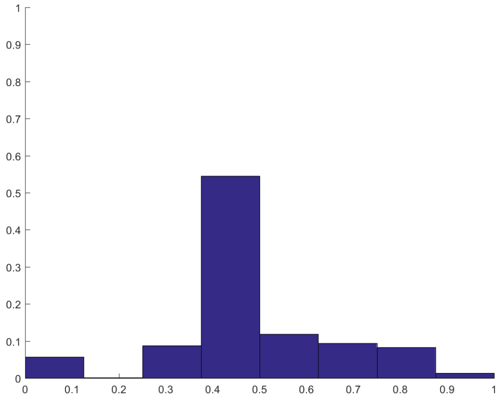
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



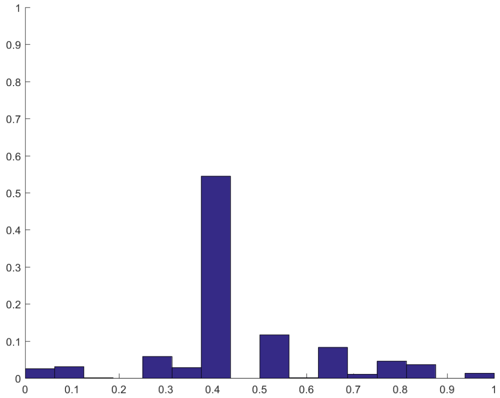
Dirichlet Process

- $\Theta = [0,1]$, G_0 uniform distribution, $\alpha = 5$



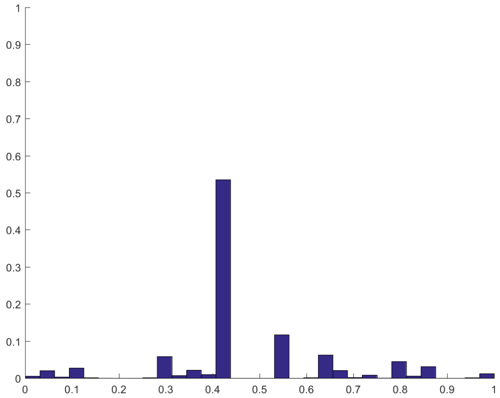
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



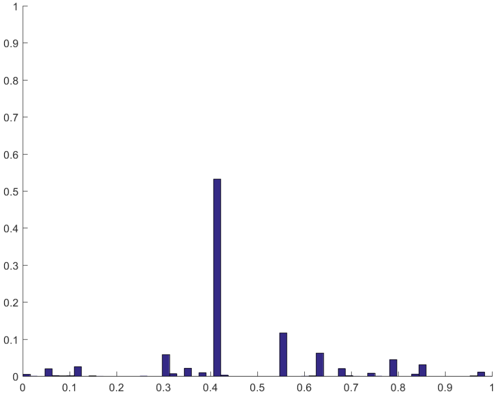
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



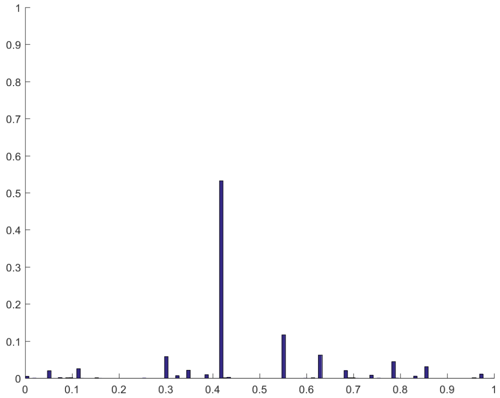
Dirichlet Process

- $\boldsymbol{\Theta = [0, 1]}$, $\boldsymbol{G_0}$ uniform distribution, $\boldsymbol{\alpha = 5}$



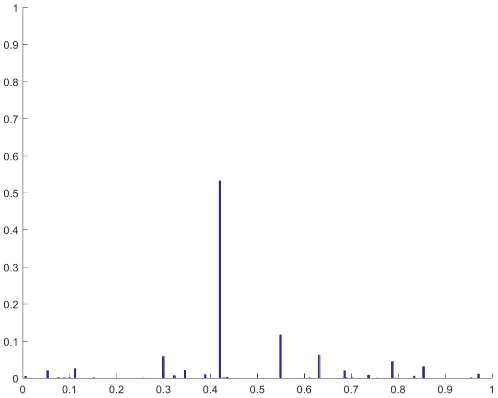
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



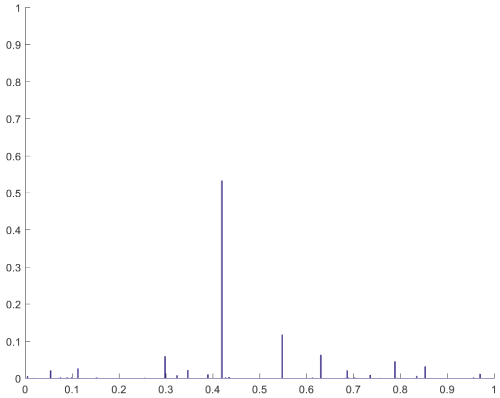
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



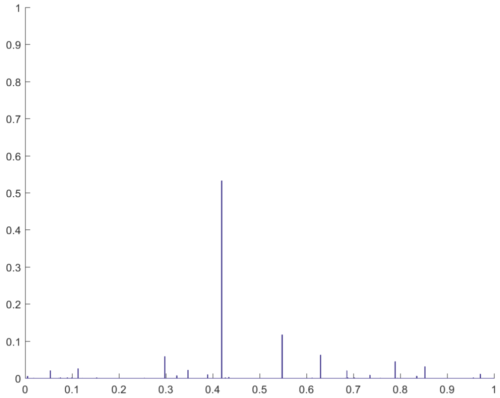
Dirichlet Process

- $\Theta = [0, 1]$, G_0 uniform distribution, $\alpha = 5$



Dirichlet Process

- $\Theta = [0,1]$, G_0 uniform distribution, $\alpha = 5$



Dirichlet Process

- From the properties of the Dirichlet distribution

$$\mathbb{E}[G(A)] = G_0(A)$$
$$\mathrm{Var}(G(A)) = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}$$

for any measurable A subset of Θ

Dirichlet Process

- Let
$$G \sim \mathrm{DP}(\alpha, G_0)$$
for $i = 1, \ldots, n$
$$\theta_i | G \overset{\text{iid}}{\sim} G$$

- Conjugacy
$$G | \theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}\right)$$

- Blackwell-MacQueen urn scheme
$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}$$

Dirichlet Process

- Proof
- Consider an arbitrary partition $A_1, \ldots, A_d$ of Θ

$$\Pr(\theta_i \in A_k | G) = G(A_k)$$

- Let $s_{n,k} = \sum_{i=1}^{n} \delta_{\theta_i}(A_k)$ be the number of θ_i falling in A_k

$$\begin{aligned}(G(A_1), ..., G(A_d))\,|\theta_{1:n} \sim \\ \mathrm{Dirichlet}(\underbrace{\alpha G_0(A_1) + s_{n,1}}_{\widetilde{\alpha}\widetilde{G}_0(A_1)}, ..., \underbrace{\alpha G_0(A_d) + s_{n,d}}_{\widetilde{\alpha}\widetilde{G}_0(A_d)})\end{aligned}$$

where $\widetilde{\alpha} = \alpha + n$ and $\widetilde{G}_0 = \frac{\alpha}{\alpha+n} G_0 + \frac{1}{\alpha+n} \sum_{i=1}^{n} \delta_{\theta_i}$.

Dirichlet Process and Chinese restaurant process

- Let $U_1, \ldots, U_{K_n}$ be the different values taken by $\theta_1, \ldots, \theta_n$ with multiplicities $m_{n,j}$
- Blackwell-MacQueen urn revisited

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha+n} G_0 + \sum_{j=1}^{K_n} \frac{m_{n,j}}{\alpha+n} \delta_{U_j}$$

- Let $\Pi_n = \{A_{n,1}, \ldots, A_{n,K_n}\}$ where $A_j = \{i|\theta_i = U_j\}$
- Then

$$\Pi_n \sim \mathrm{CRP}(\alpha, n)$$

and

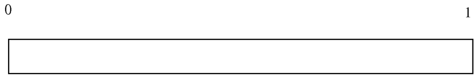
$$U_j \overset{\text{iid}}{\sim} G_0$$

Dirichlet Process

- Realization of a DP is a.s. discrete and admits the following *stick-breaking* representation

$$G = \sum_{j=1}^{\infty} \pi_j \delta_{U_j}$$

with $\pi_j = \beta_j \prod_{k<j}(1-\beta_k)$, $\beta_j \sim \mathrm{Beta}(1,\alpha)$ and $U_j \overset{\text{iid}}{\sim} G_0$.



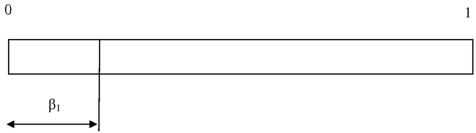
[Sethuraman, 1994]

Dirichlet Process

- Realization of a DP is a.s. discrete and admits the following *stick-breaking* representation

$$G = \sum_{j=1}^{\infty} \pi_j \delta_{U_j}$$

with $\pi_j = \beta_j \prod_{k<j}(1-\beta_k)$, $\beta_j \sim \text{Beta}(1,\alpha)$ and $U_j \overset{\text{iid}}{\sim} G_0$.



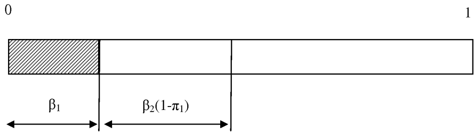
[Sethuraman, 1994]

Dirichlet Process

- Realization of a DP is a.s. discrete and admits the following *stick-breaking* representation

$$G = \sum_{j=1}^{\infty} \pi_j \delta_{U_j}$$

with $\pi_j = \beta_j \prod_{k<j}(1-\beta_k)$, $\beta_j \sim \text{Beta}(1,\alpha)$ and $U_j \overset{\text{iid}}{\sim} G_0$.



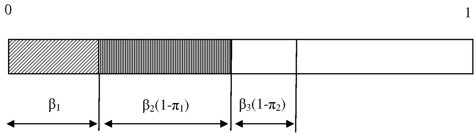
[Sethuraman, 1994]

Dirichlet Process

- Realization of a DP is a.s. discrete and admits the following *stick-breaking* representation

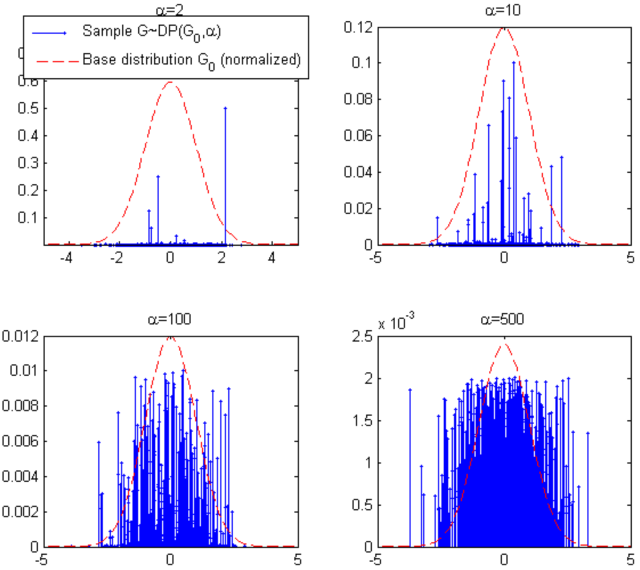
$$G = \sum_{j=1}^{\infty} \pi_j \delta_{U_j}$$

with $\pi_j = \beta_j \prod_{k<j}(1-\beta_k)$, $\beta_j \sim \mathrm{Beta}(1,\alpha)$ and $U_j \overset{\text{iid}}{\sim} G_0$.



[Sethuraman, 1994]

Dirichlet Process



Dirichlet Process Mixture

- The data $\mathbf{y}_i$ are supposed to be distributed from the following mixture model

$$\mathbf{y}_i | G \overset{\text{iid}}{\sim} \int_{\Theta} f_U(\cdot) G(dU)$$

where the mixing distribution G is unknown

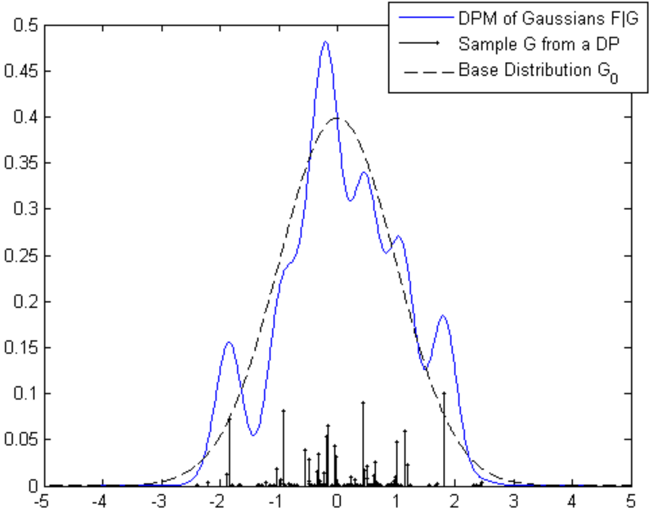
$$G \sim \mathrm{DP}(\alpha, G_0)$$

- Using the stick-breaking representation

$$\int_{\Theta} f_U(\cdot) G(dU) = \sum_{j=1}^{\infty} \pi_j f_{U_j}(\cdot)$$

- Infinite mixture model

Dirichlet Process Mixture



Dirichlet Process Mixture

- Hierarchical model

$$G \sim \mathrm{DP}(\alpha, G_0)$$

for $i = 1, \ldots, n$

$$\theta_i | G \sim G$$
$$\mathbf{y}_i | \theta_i \sim f_{\theta_i}$$

- This model is equivalent to

$$\Pi_n \sim \mathrm{CRP}(\alpha, n)$$

for $j = 1, \ldots, K_n$,

$$U_j \sim G_0$$

For $i = 1, \ldots, n$

$$y_i | \Pi_n, U_1, \ldots, U_{K_n} \sim f_{U_{c_i}}$$

Slice sampling for Dirichlet Process Mixtures

- The previous sampler was a marginalized sampler, as $\boldsymbol{G}$ is marginalized out
- One drawback: does not scale well with the number of data (no parallelization possible)
- Hierarchical sampler: full posterior $\boldsymbol{p(G, c_{1:n}|y_{1:n})}$

Slice sampling

- Suppose we want to sample from a distribution $f(x)/Z$ where $Z = \int f(x)dx$.
- Introduce a latent slice variable $u > 0$
- Joint distribution

$$p(x,u) = \begin{cases} 1/Z & \text{if } 0 < u < f(x) \\ 0 & \text{otherwise} \end{cases}$$

- Marginal distribution over x

$$p(x) = \int p(x,u)du = \int_0^{f(x)} \frac{1}{Z} = \frac{f(x)}{Z}$$

Neal, 2003

Slice sampling

- Slice sampling: MCMC algorithm with target distribution $p(x, u)$
- At each iteration
  - Sample $u|x \sim \mathrm{Unif}([0, f(x)])$
  - Sample $x|u \sim \mathrm{Unif}(\{x|f(x) > u\})$
- Example: We want to sample from the discrete distribution $G = \sum_{j=1}^{\infty} \pi_j \delta_{U_j}$
- At each iteration
  - Sample $u|x = U_j \sim \mathrm{Unif}([0, \pi_j])$
  - Sample $x|u \sim \mathrm{Unif}(\{U_j|\pi_j > u\})$

Slice sampling for Dirichlet process mixtures

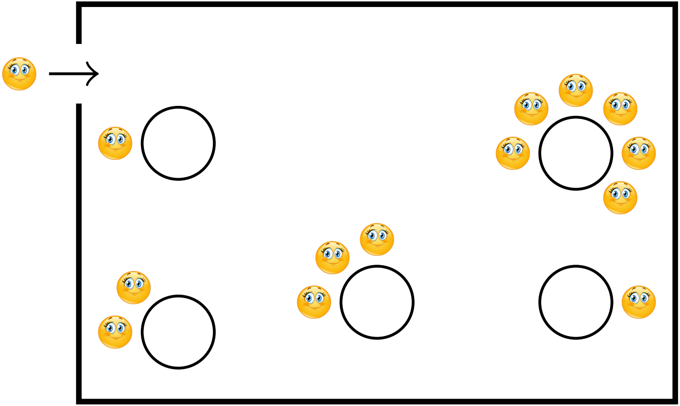
- Latent slice variables u_k, $k = 1, \ldots, n$
- Let m_j be the number of allocation variables taking value $j \in \{1, \ldots, K\}$
- At each iteration
  - Sample $(\pi_1, \ldots, \pi_K, \pi_*) \sim \text{Dirichlet}(m_1, \ldots, m_K, \alpha)$
  - For $k = 1, \ldots, n$ sample $u_k \sim \mathbf{Unif}([0, \pi_{c_k}])$
  - Set $\ell = K$. While $\sum_{j=1}^{\ell} \pi_j < (1 - \min(u_1, \ldots, u_n))$
    - Set $\ell = \ell + 1$
    - Sample $\beta_\ell \sim \text{Beta}(1, \alpha)$
    - Set $\pi_\ell = \pi_* \beta_\ell \prod_{j=K+1}^{\ell-1} (1 - \beta_j)$
    - Sample $U_\ell \sim G_0$
  - For $i = 1, \ldots, n$ sample c_i from

$$p(c_i = j) \propto \mathbf{1}(\pi_j > u_i) f(y_i | U_j)$$

  - For $j = 1, \ldots, K$ sample $U_j |$rest

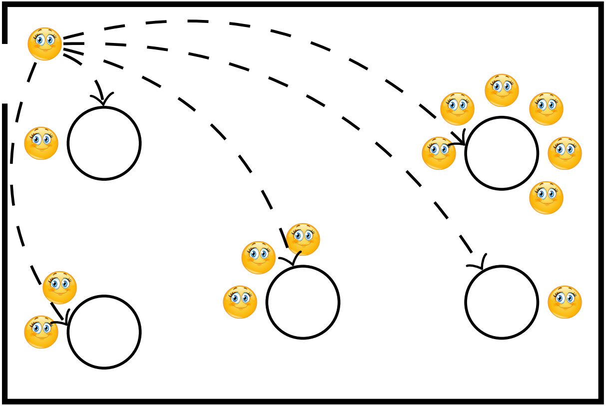
Walker (2007), Kalli et al. (2011), Fall and Barat (2012)

Two-parameter Chinese restaurant process



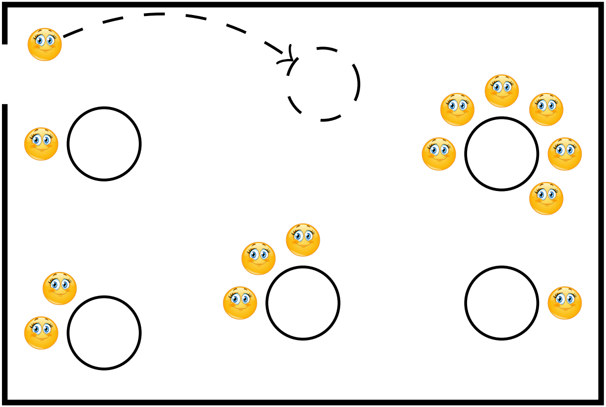
- Customer $n+1$
  - Joins an existing table $k=1,\ldots,K_n$ w.p. $\frac{m_{n,k}-\sigma}{n+\alpha}$
  - Sits at a new table w.p. $\frac{K_n\sigma+\alpha}{n+\alpha}$

Two-parameter Chinese restaurant process



- Customer $n+1$
  - Joins an existing table $k=1,\ldots,K_n$ w.p. $\frac{m_{n,k}-\sigma}{n+\alpha}$
  - Sits at a new table w.p. $\frac{K_n\sigma+\alpha}{n+\alpha}$

Two-parameter Chinese restaurant process



- Customer $n+1$
  - Joins an existing table $k=1,\dots,K_n$ w.p. $\frac{m_{n,k}-\sigma}{n+\alpha}$
  - Sits at a new table w.p. $\frac{K_n\sigma+\alpha}{n+\alpha}$

Two-parameter Chinese restaurant process

- Rich-gets-richer process

$$\boldsymbol{\Pi_n \sim \mathrm{CRP}(\sigma, \alpha, n)}$$

- Two parameters $0 \leq \sigma < 1$, $\alpha > -\sigma$
- $\sigma = 0$: One-parameter CRP
- Exchangeable random partition
- Growth of the number of clusters

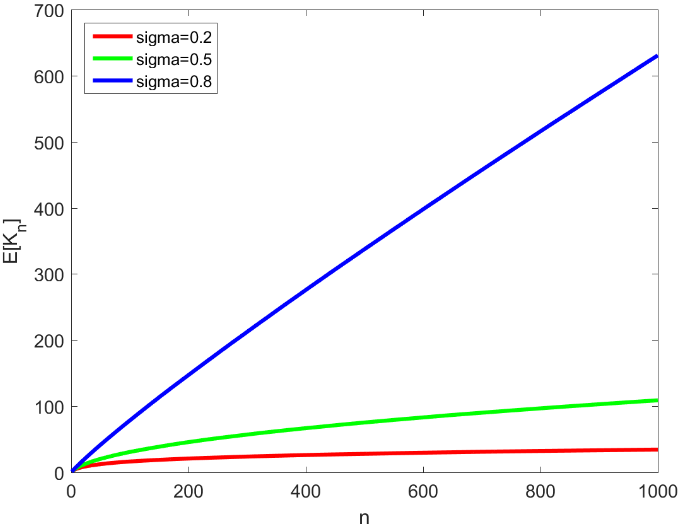
$$K_n = \begin{cases} \Theta(\log n) & \text{if } \sigma = 0 \\ \Theta(n^\sigma) & \text{if } \sigma > 0 \end{cases} \qquad \text{a.s. as } n \to \infty$$

- Power-law behavior for $\sigma > 0$
  - Let $K_{n,j}$ be the number of clusters of size j

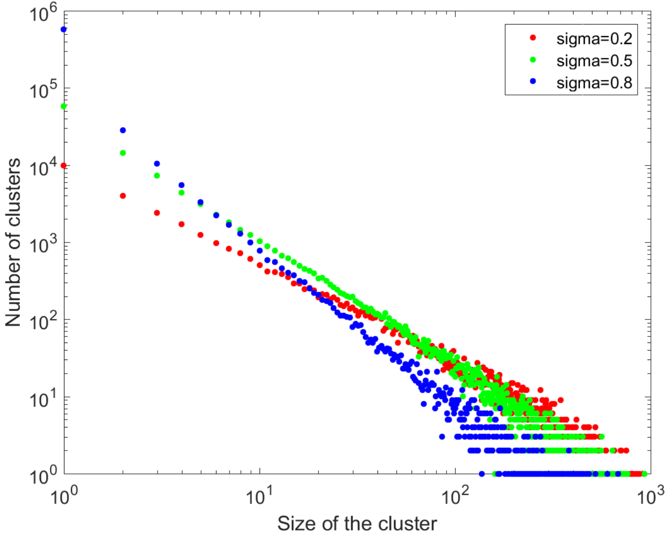
$$\frac{K_{n,j}}{K_n} \to p_j \text{ almost surely as } n \to \infty$$

    where p_j is of order $j^{-1-\sigma}$
- Various applications in natural language or image processing

Two-parameter Chinese restaurant process



Two-parameter Chinese restaurant process



Outline

Introduction

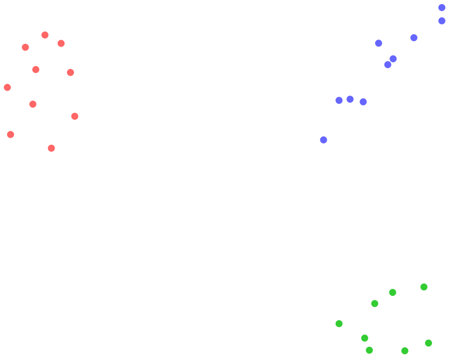
Clustering

- Cluster/partition a set of items $i = 1, \ldots, n$ into clusters

Introduction

Clustering

- Cluster/partition a set of items $i = 1, \ldots, n$ into clusters



- Random partition

$$\Pi_n = \{A_{n,1}, \ldots, A_{n,K_n}\}$$

where $A_{n,j}$, $j = 1, \ldots, K_n$ non-empty and non-overlapping subsets of $[n] := \{1, \ldots, n\}$ with $\cup_{j=1}^{K_n} A_{n,j} = [n]$

- $A_{n,j}$ are clusters, $K_n \leq n$ is the number of clusters
- Example

$$\Pi_6 = \{\{1, 4, 5\}, \{2, 3\}, \{6\}\}$$

Introduction
Clustering

- Nonparametric approach: K_n can increase unboundedly with the number of items n
- Exchangeable random partition: Distribution is invariant w.r.t. any permutation of $[n]$, e.g.

$$P(\{\{1,2\},\{3\}\}) = P(\{\{2,3\},\{1\}\}) = P(\{\{1,3\},\{2\}\})$$

- Labelling/ordering of the items is of no importance
- Chinese restaurant process is an example of a generative process for an exchangeable partition

Introduction

Latent feature models

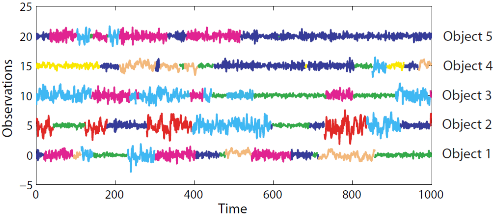
- Set of objects $i = 1, \dots, n$
- Objects i have a set of features/attributes, shared amongst objects
- Example:

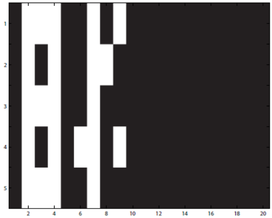
Image 1				
Image 2	Tree	Human		
Image 3		Human		
Image 4	Tree	Human		
Image 5			Road	Animal

Introduction

Latent feature models

- Dynamic state-space models
- Collection of time series with shared dynamical behaviors



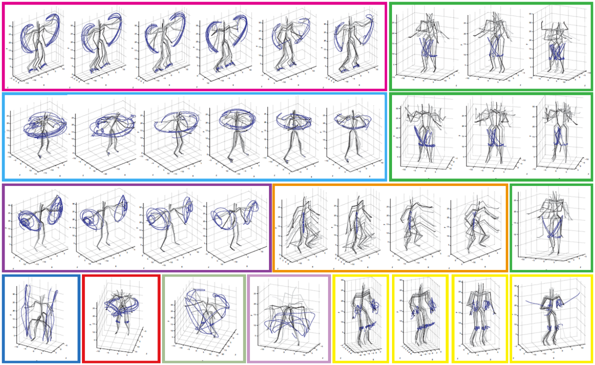


[Fox et al., 2009]

Introduction

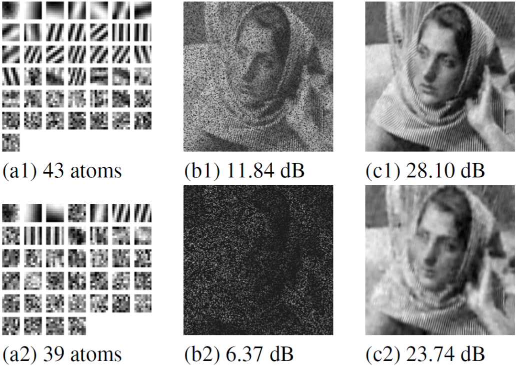
Latent feature models

- Application to dynamic state-space models
- Collection of time series with shared dynamical behaviors



[Fox et al., 2009]

- Dictionary learning for image inpainting

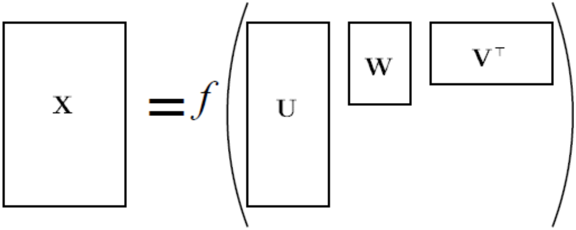


[Zhou et al., 2009, Dang and Chainais, 2016]

Introduction

Latent feature models

- Collaborative filtering: predict missing entries in a user/items matrix from a subset of its entries
- Low-rank assumption: matrix can be decomposed with a small number of latent features
- User/feature association matrix



[Meeds et al., 2007]

Introduction

Latent feature models

- Random feature allocation
- Representation as a multiset of $[n] = \{1, \dots, n\}$

$$f_n = \{A_{n,1}, \dots, A_{n,K_n}\}$$

where $A_{n,j}$, $j = 1, \dots, K_n$ are non-empty (possibly overlapping) subsets of $[n]$

- $A_{n,j}$, $j = 1, \dots, K_n$ are sets of objects sharing a given feature j
- Example:

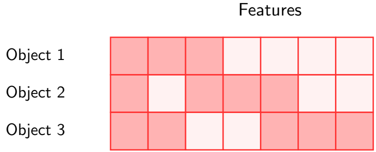
$$f_5 = \{\{2, 3, 4\}, \{2, 4\}, \{5\}, \{5\}\}$$

Image 1				
Image 2	Tree	Human		
Image 3		Human		
Image 4	Tree	Human		
Image 5			Road	Animal

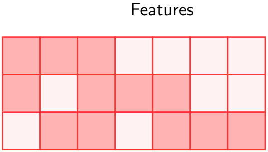
[Broderick et al., 2013a]

Introduction
Latent feature models

- Multisets often graphically represented by a binary matrix
- Beware that feature labelling does not matter!



and



represent the same multiset

$$f_3 = \{\{1,2,3\},\{1,3\},\{1,2\},\{2\},\{2,3\},\{3\},\{3\}\}$$

Introduction

Latent feature models

- Nonparametric approach: the number of features K_n can increase unboundedly with n
- Exchangeable latent feature model: distribution of f_n invariant w.r.t. any permutation σ of $[n]$, e.g.

$$
\begin{aligned}
&\Pr(\{\{2,3,4\},\{2,4\},\{5\},\{5\}\}) \\
&= \Pr(\{\{3,4,5\},\{3,5\},\{1\},\{1\}\}) \\
&= \Pr(\{\{\sigma(2),\sigma(3),\sigma(4)\},\{\sigma(2),\sigma(4)\},\{\sigma(5)\},\{\sigma(5)\}\})
\end{aligned}
$$

for any permutation σ of $\{1,2,3,4,5\}$

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \text{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \text{Poisson}(\alpha/i)$

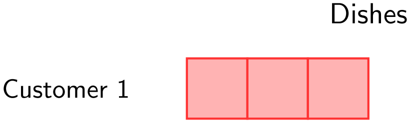
Dishes

Customer 1

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \text{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \text{Poisson}(\alpha/i)$



$$f_1 = \{\{1\}, \{1\}, \{1\}\}$$

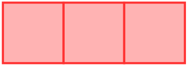
[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \textbf{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \textbf{Poisson}(\alpha/i)$

Dishes

Customer 1

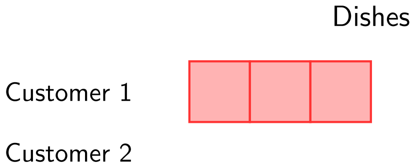


$$f_1 = \{\{1\}, \{1\}, \{1\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \mathrm{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \mathrm{Poisson}(\alpha/i)$



$$f_1 = \{\{1\}, \{1\}, \{1\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

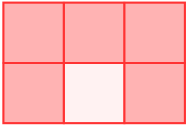
Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \textbf{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \textbf{Poisson}(\alpha/i)$

Dishes

Customer 1

Customer 2

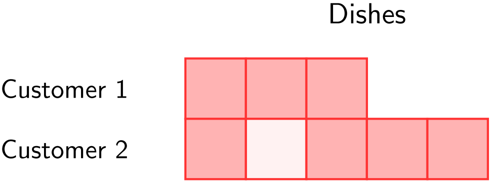


$$\{\{1, 2\}, \{1\}, \{1, 2\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \textbf{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \textbf{Poisson}(\alpha/i)$

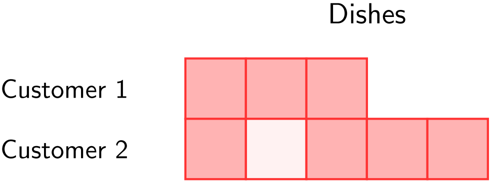


$$f_2 = \{\{1, 2\}, \{1\}, \{1, 2\}, \{2\}, \{2\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \textbf{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \textbf{Poisson}(\alpha/i)$

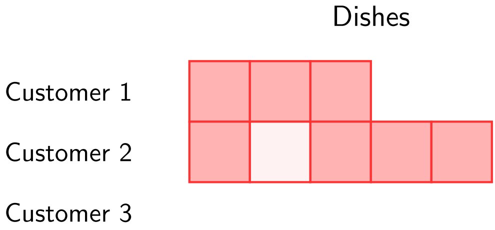


$$f_2 = \{\{1, 2\}, \{1\}, \{1, 2\}, \{2\}, \{2\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \mathbf{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \mathbf{Poisson}(\alpha/i)$

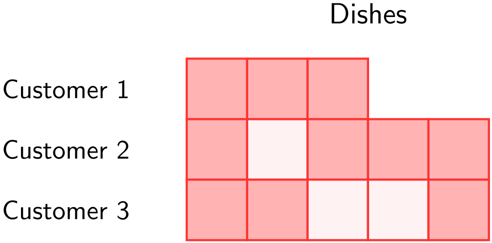


$$f_2 = \{\{1,2\},\{1\},\{1,2\},\{2\},\{2\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \text{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \text{Poisson}(\alpha/i)$

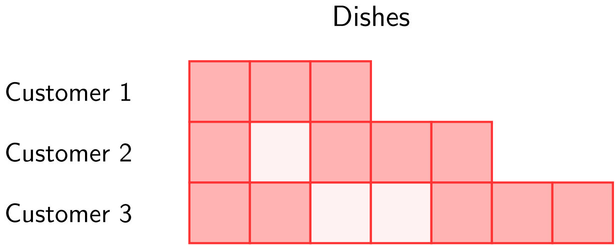


$$\{\{1,2,3\},\{1,3\},\{1,2\},\{2\},\{2,3\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process

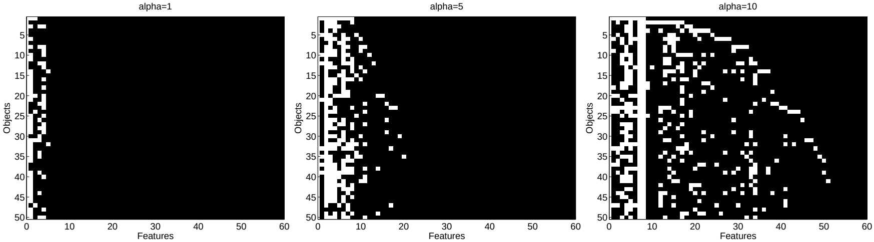
- Generative model for multisets
- Single parameter $\alpha > 0$
- First customer picks $K_1^+ \sim \text{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability $m_{i-1,j}/i$
  - picks an additional set of dishes $K_i^+ \sim \text{Poisson}(\alpha/i)$



$$f_3 = \{\{1, 2, 3\}, \{1, 3\}, \{1, 2\}, \{2\}, \{2, 3\}, \{3\}, \{3\}\}$$

[Griffiths and Ghahramani, 2005, Griffiths and Ghahramani, 2011]

Indian buffet process



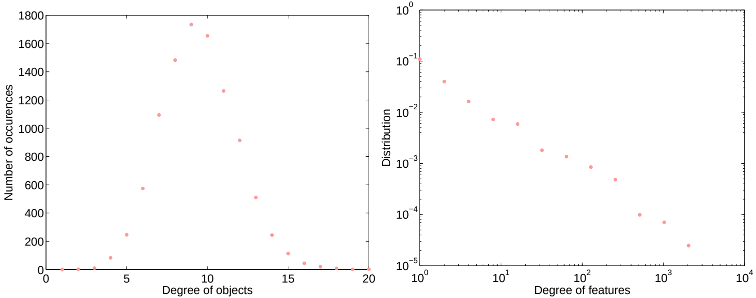
Indian buffet process

- Rich gets richer process: more popular dishes are more likely to be chosen by new customers
- New dishes can always be picked as new customers arrive, but at a decreasing rate α/i
- Number of features/dishes for n customers follows a $\textbf{Poisson}$ distribution with rate

$$\alpha \sum_{i=1}^{n} \frac{1}{i} \simeq \alpha \log(n)$$

- Number of dishes picked by each customer (degree of a customer) follows $\textbf{Poisson}(\alpha)$
- Degree distribution of features follows a heavy tail distribution

Indian buffet process



Indian buffet process

- Multiset $f_n = \{A_{n,1}, \ldots, A_{n,K_n}\}$ with $m_{n,j} = |A_{n,j}|$
- Let $\{\widetilde{A}_{n,1}, \ldots, \widetilde{A}_{n,\widetilde{K}_n}\}$ be the set of unique values in f_n, and $\kappa_1, \ldots, \kappa_{\widetilde{K}_n}$ be their multiplicities, then

$$\Pr(f_n) = \frac{\alpha^{K_n}}{\prod_{h=1}^{\widetilde{K}_n} \kappa_h!} e^{-\alpha \sum_{i=1}^{n} \frac{1}{i}} \prod_{j=1}^{K_n} \frac{(m_{n,j}-1)!(n-m_{n,j})!}{n!}$$

- Does not depend on the ordering of the customers
- Exchangeable latent feature model

Indian buffet process

- How to derive the IBP?
  - Limit of a parametric beta Bernoulli model
  - Completely random measures

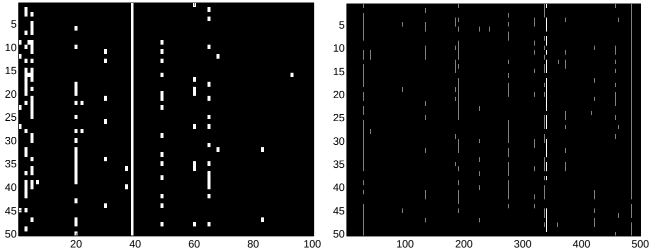
Parametric beta Bernoulli model

- Binary matrix $\boldsymbol{z} = (z_{i,j})$ of size $n \times p$
- For $j = 1, \ldots, p$

$$\pi_j \sim \text{Beta}\left(\frac{\alpha}{p}, 1\right)$$

- For $i = 1, \ldots, n$ and $j = 1, \ldots, p$

$$z_{i,j} | \pi_j \sim \text{Ber}(\pi_j)$$



(a) $p = 100$ (b) $p = 500$

Parametric beta Bernoulli model

$$\Pr(z) = \prod_{j=1}^{p} \int_0^1 \prod_{i=1}^{n} \pi_j^{z_{i,j}} (1 - \pi_j)^{1 - z_{i,j}} \, \text{Beta}(\pi_j; \alpha/p, 1) d\pi_j$$

Parametric beta Bernoulli model

$$\begin{aligned}
\Pr(z) &= \prod_{j=1}^{p} \int_0^1 \prod_{i=1}^{n} \pi_j^{z_{i,j}} (1-\pi_j)^{1-z_{i,j}} \, \text{Beta}(\pi_j; \alpha/p, 1) d\pi_j \\
&= \prod_{j=1}^{p} \int_0^1 \pi_j^{\sum_i z_{ij}} (1-\pi_j)^{n-\sum_i z_{ij}} \, \text{Beta}(\pi_j; \alpha/p, 1) d\pi_j
\end{aligned}$$

Parametric beta Bernoulli model

$$
\begin{aligned}
\Pr(z) &= \prod_{j=1}^{p} \int_0^1 \prod_{i=1}^{n} \pi_j^{z_{i,j}} (1-\pi_j)^{1-z_{i,j}} \,\mathrm{Beta}(\pi_j; \alpha/p, 1) d\pi_j \\
&= \prod_{j=1}^{p} \int_0^1 \pi_j^{\sum_i z_{ij}} (1-\pi_j)^{n-\sum_i z_{ij}} \,\mathrm{Beta}(\pi_j; \alpha/p, 1) d\pi_j \\
&= \prod_{j=1}^{p} \frac{B(\sum_i z_{ij} + \alpha/p, n - \sum_i z_{ij} + 1)}{B(\alpha/p, 1)}
\end{aligned}
$$

Parametric beta Bernoulli model

$$\begin{aligned}
\Pr(z) &= \prod_{j=1}^{p} \int_0^1 \prod_{i=1}^{n} \pi_j^{z_{i,j}} (1-\pi_j)^{1-z_{i,j}} \,\mathrm{Beta}(\pi_j;\alpha/p,1) d\pi_j \\
&= \prod_{j=1}^{p} \int_0^1 \pi_j^{\sum_i z_{ij}} (1-\pi_j)^{n-\sum_i z_{ij}} \,\mathrm{Beta}(\pi_j;\alpha/p,1) d\pi_j \\
&= \prod_{j=1}^{p} \frac{B(\sum_i z_{ij}+\alpha/p, n-\sum_i z_{ij}+1)}{B(\alpha/p,1)} \\
&= \prod_{j=1}^{p} \frac{\alpha/p\Gamma(\sum_i z_{ij}+\alpha/p)\Gamma(n-\sum_i z_{ij}+1)}{\Gamma(n+1+\alpha/p)}
\end{aligned}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function, using $\Gamma(a+1) = a\Gamma(a)$.

Parametric beta Bernoulli model

- Let $f_n = \text{multiset}(z)$ denote the multiset corresponding to z

$$\text{multiset}(z) = \{\{i|z_{ij} = 1\}, j = 1, \ldots, p \text{ s.t.} \sum_i z_{ij} > 0\}$$

- Many matrices z correspond to the same multiset
- Let $E(f_n) = \{z|f_n = \text{multiset}(z)\}$ be the set of matrices corresponding to the same multiset f_n
- Cardinality of $E(f_n)$

$$|E(f_n)| = \frac{p!}{\kappa_0! \prod_{h=1}^{\widetilde{K}_n} \kappa_h!}$$

where κ_0 is the number of all-zero columns.

Parametric beta Bernoulli model

- Due to column exchangeability, all matrices $z \in E(f_n)$ have the same probability

$$
\begin{aligned}
\Pr(f_n) &= \sum_{z \in E(f_n)} \Pr(z) \\
&= \frac{p!}{\kappa_0! \prod_{h=1}^{\widetilde{K}_n} \kappa_h!} \prod_{j=1}^{K_n} \frac{\alpha/p \Gamma(m_{n,j} + \alpha/p) \Gamma(n - m_{n,j} + 1)}{\Gamma(n + 1 + \alpha/p)} \\
&\qquad \times \left(\frac{\alpha/p \Gamma(\alpha/p) \Gamma(n+1)}{\Gamma(n+1+\alpha/p)} \right)^{\kappa_0} \\
&= \frac{\alpha^{K_n}}{\prod_{h=1}^{\widetilde{K}_h} \kappa_h!} \frac{p!}{\kappa_0! p^{K_n}} \left(\frac{n! \Gamma(\alpha/p + 1)}{\Gamma(n+1+\alpha/p)} \right)^p \\
&\times \prod_{j=1}^{K_n} \frac{\Gamma(m_{n,j} + \alpha/p)(n - m_{n,j})!}{\Gamma(\alpha/p + 1) n!}
\end{aligned}
$$

Parametric beta Bernoulli model

- Taking the limit as $p \to \infty$

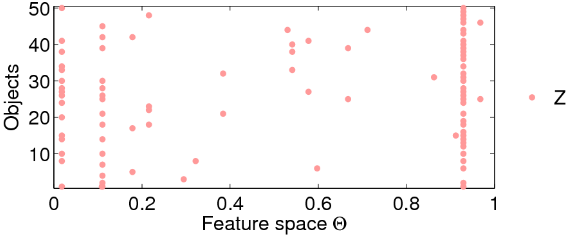
$$\frac{\alpha^{K_n}}{\prod_{h=1}^{\widetilde{K}_h} \kappa_h!} \frac{p!}{\kappa_0! p^{K_n}} \left(\frac{n!\Gamma(\alpha/p+1)}{\Gamma(n+1+\alpha/p)}\right)^p \times \prod_{j=1}^{K_n} \frac{\Gamma(m_{n,j}+\alpha/p)(n-m_{n,j})!}{\Gamma(\alpha/p+1)n!} \overrightarrow{p \to \infty} \frac{\alpha^{K_n}}{\prod_{h=1}^{\widetilde{K}_h} \kappa_h!} \cdot 1 \cdot e^{-\alpha \sum_{i=1}^{n} 1/i} \times \prod_{j=1}^{K_n} \frac{(m_{n,j}-1)!(n-m_{n,j})!}{n!}$$

Beta-Bernoulli process

- Now assume that each feature $j = 1, \dots, K_n$ has some location $\theta^\star_{n,j}$ in a feature space Θ
- Feature locations are assumed to be i.i.d from some distribution G_0 (density g_0)
- Represent the feature model as a collection of point processes

$$Z_i = \sum_{j=1}^{\infty} z_{ij} \delta_{\theta_j}$$

where δ_a is the dirac delta mass and
  - $z_{ij} = 1$ if object i possesses feature θ_j
  - $\{\theta^\star_{n,j}\} = \{\theta_k | \exists i \in [n] \; s.t. \; z_{ik} > 0\}$



Beta-Bernoulli process

- Let $f_n(Z_1,\dots,Z_n)$ be the multiset induced by the point processes

$$f_n(Z_1,\dots,Z_n)=\{\{i|Z_i(\theta^\star_{n,j})=1\},j=1,\dots,K_n\}$$

- Distribution over $(Z_i)_{i=1,\dots,n}$ is obtained by setting independent priors over the feature allocations and their locations

$$p(Z_1,\dots,Z_n)=\Pr(f_n(Z_1,\dots,Z_n))\prod_{j=1}^{K_n}g_0(\theta^\star_{n,j})\prod_{h=1}^{\widetilde{K}_h}\kappa_h!$$

- Using the IBP prior for the feature allocations

$$p(Z_1,\dots,Z_n)=\alpha^{K_n}e^{-\alpha\sum_{i=1}^n\frac{1}{i}}\prod_{j=1}^{K_n}\frac{(m_{n,j}-1)!(n-m_{n,j})!}{n!}$$

$$\times\prod_{j=1}^{K_n}g_0(\theta^\star_j)$$

Beta-Bernoulli process

- Exchangeability over the feature allocations f_n carries over $(Z_i)_{i=1,\dots,n}$
- Infinite exchangeability: for any $n \geq 1$ and any permutation σ of $[n]$

$$p(Z_1, \dots, Z_n) = p(Z_{\sigma(1)}, \dots, Z_{\sigma(n)})$$

- De Finetti representation theorem implies

$$p(Z_1, \dots, Z_n) = \int \prod_{i=1}^{n} p(Z_i | B) P(dB)$$

where B is some latent process with distribution P

- de Finetti measure $P(dB)$: beta process

[Hjort, 1990, Thibaux and Jordan, 2007]

Beta-Bernoulli process

- Let

$$B = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

be a completely random measure characterized by its Lévy measure

$$\nu(d\pi, d\theta) = \alpha \pi^{-1}(1-\pi)^{\alpha-1} d\pi G_0(d\theta)$$

defined on $[0,1] \times \Theta$.

- B is called a beta process and we write

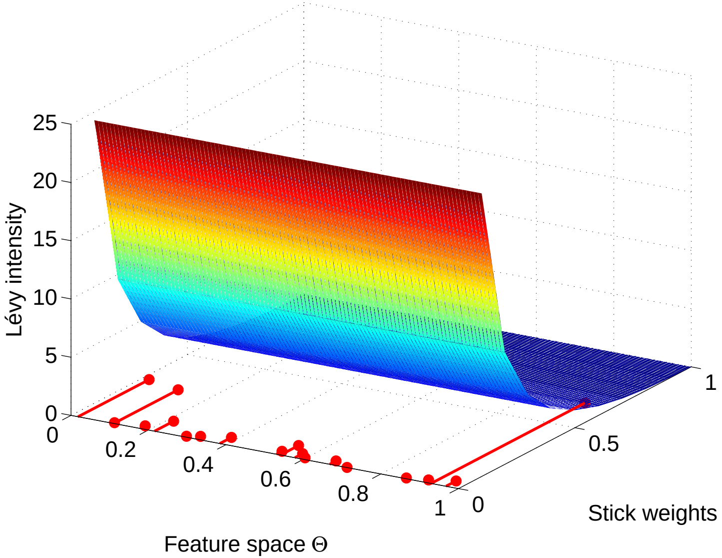
$$B \sim \mathrm{BetaP}(\alpha, G_0)$$

- A draw from a beta process is discrete a.s. with an infinite number of atoms

[Hjort, 1990]

Beta-Bernoulli process

- Beta process



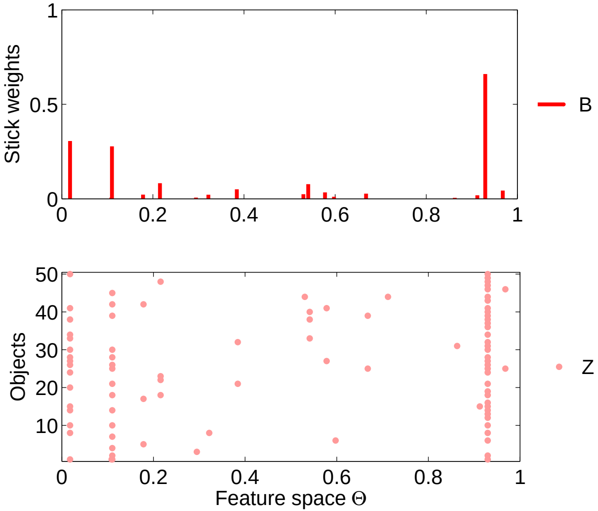
Beta-Bernoulli process

- Conditional Bernoulli process

$$Z_i | B \sim \mathrm{BeP}(B)$$

$$Z_i = \sum_{j=1}^{\infty} z_{ij} \delta_{\theta_j} \text{ where } z_{ij} \sim \mathrm{Ber}(\pi_j)$$

Beta-Bernoulli process



Beta-Bernoulli process

- Conjugacy
- Let $\theta^\star_{n,1}, \ldots, \theta^\star_{n,K_n}$ be the number of support points in $Z_1, \ldots, Z_n$ and $m_{n,j}$ their occurences
- Posterior

$$B|Z_1, \ldots, Z_n \sim \mathrm{BetaP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \sum_{j=1}^{K_n} \frac{m_{n,j}}{\alpha + n} \delta_{\theta^\star_{n,j}}\right)$$

- Predictive distribution

$$Z_{n+1}|Z_1, \ldots, Z_n \sim \mathrm{BeP}\left(\frac{\alpha}{\alpha + n} G_0 + \sum_{j=1}^{K_n} \frac{m_{n,j}}{\alpha + n} \delta_{\theta^\star_{n,j}}\right)$$

[Hjort, 1990, Kim, 1999, Thibaux and Jordan, 2007]

Chinese restaurant vs Indian buffet

Application	Clustering	Latent feature
Combinatorial object	Partition	Multiset
Generative model	Chinese restaurant proc.	Indian buffet proc.
de Finetti measure	Dirichlet process	beta process
Stick-breaking	Yes	Yes
Conjugacy	Yes	Yes
Power-law extensions	Pitman-Yor	stable beta process

Inference

- Latent variable model
- Data $\boldsymbol{X}$ of size $\boldsymbol{n \times d}$
- (Marginal) Likelihood

$$\mathbf{Pr}(\boldsymbol{X}|\boldsymbol{f_n}) = \int_{\boldsymbol{\Theta}} \mathbf{Pr}(\boldsymbol{X}|\boldsymbol{f_n}, \boldsymbol{\theta}) \boldsymbol{P}(\boldsymbol{\theta}) \boldsymbol{d\theta}$$

- Prior

$$\mathbf{Pr}(\boldsymbol{f_n})$$

- Posterior

$$\mathbf{Pr}(\boldsymbol{f_n}|\boldsymbol{X}) \propto \mathbf{Pr}(\boldsymbol{X}|\boldsymbol{f_n})\,\mathbf{Pr}(\boldsymbol{f_n})$$

- Inference can be carried out using IBP
  - MCMC with Metropolis-Hastings within Gibbs updates
  - Sequential Monte Carlo

[Meeds et al., 2007, Wood and Griffiths, 2007]

Stable Indian buffet process

- Three parameters $\alpha > 0$, $\sigma \in [0, 1)$ and $c > -\sigma$
- First customer picks $K_1^+ \sim \text{Poisson}(\alpha)$ dishes
- Then each customer $i = 2, \ldots$
  - chooses a dish j previously chosen $m_{i-1,j}$ times with probability

$$\frac{m_{i-1,j} - \sigma}{c + i - 1}$$

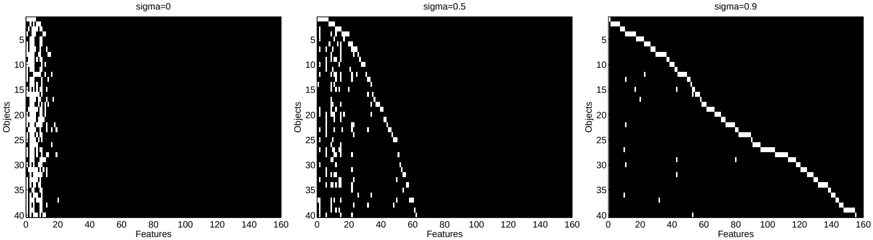
  - picks an additional set of dishes

$$K_i^+ \sim \text{Poisson}\left(\alpha \frac{\Gamma(1+c)\Gamma(i-1+c+\sigma)}{\Gamma(i+c)\Gamma(c+\sigma)}\right)$$

- Reduces to the one parameter IBP when $c = 1$ and $\sigma = 0$

[Teh and Görür, 2009]

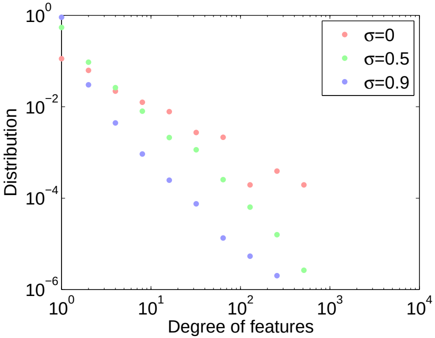
Stable Indian buffet process

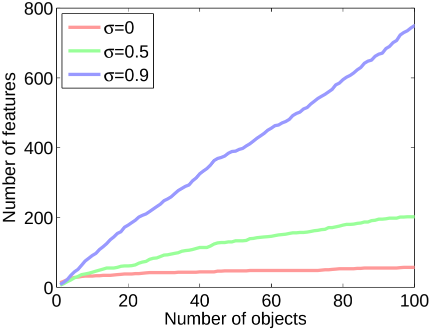


Stable Indian buffet process

- Power-law behavior for $\sigma > 0$
- Number of features grows in $O(n^{\sigma})$
- Proportion of features associated to m objects is, for $n \gg m$ large, in $O\left(\frac{1}{m^{1+\sigma}}\right)$
- Similar to the Pitman-Yor process for mixture models

Stable Indian buffet process





Outline

Conclusion

- Bayesian nonparametrics offers a robust and adaptive framework
- Mathematically more involved, but inference algorithms are often as simple as the parametric ones
- Many other models and applications of BNP
- Standard tools for Bayesian modeling

Bibliography I

Bar-Hillel, A., Spiro, A., and Stark, E. (2006).
Spike sorting: Bayesian clustering of non-stationary data.
Journal of neuroscience methods, 157(2):303–316.

Blei, D. M. (2012).
Probabilistic topic models.
Communications of the ACM, 55(4):77–84.

Broderick, T., Jordan, M. I., and Pitman, J. (2013a).
Cluster and feature modeling from combinatorial stochastic processes.
Statistical Science, 28(3):289–312.

Broderick, T., Pitman, J., and Jordan, M. I. (2013b).
Feature allocations, probability functions, and paintboxes.
Bayesian Analysis, 8(4):801–836.

Caron, F., Neiswanger, W., Wood, F., Doucet, A., and Davy, M. (2016).
Generalized Pólya urn for time-varying pitman-yor processes.
to appear in Journal of Machine Learning Research.

Dang, H. P. and Chainais, P. (2016).
Indian buffet process dictionary learning for image inpainting.
In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5.

Bibliography II

Escobar, M. D. and West, M. (1995).
Bayesian density estimation and inference using mixtures.
Journal of the american statistical association, 90(430):577–588.

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
The annals of statistics, pages 209–230.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2009).
Sharing features among dynamical systems with beta processes.
In *Advances in Neural Information Processing Systems*, volume 22, pages 549–557.

Gasthaus, J., Wood, F., Görür, D., and Teh, Y. W. (2008).
Dependent Dirichlet process spike sorting.
In *NIPS*, pages 497–504.

Griffiths, T. and Ghahramani, Z. (2005).
Infinite latent feature models and the Indian buffet process.
In *NIPS*.

Griffiths, T. and Ghahramani, Z. (2011).
The Indian buffet process: an introduction and review.
Journal of Machine Learning Research, 12(April):1185–1224.

Bibliography III

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

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010).
Bayesian nonparametrics, volume 28.
Cambridge University Press.

Kim, Y. (1999).
Nonparametric Bayesian estimators for counting processes.
Annals of Statistics, pages 562–588.

Lijoi, A. and Prünster, I. (2010).
Models beyond the Dirichlet process.
In N. L. Hjort, C. Holmes, P. M. S. G. W., editor, *Bayesian Nonparametrics*. Cambridge University Press.

Meeds, E., Ghahramani, Z., Neal, R., and Roweis, S. (2007).
Modeling dyadic data with binary latent factors.
In *NIPS*, volume 19, page 977. MIT; 1998.

Müller, P. and Quintana, F. A. (2004).
Nonparametric Bayesian data analysis.
Statistical science, pages 95–110.

Bibliography IV

Neal, R. (2000).
Markov chain sampling methods for Dirichlet process mixture models.
Journal of computational and graphical statistics, pages 249–265.

Pitman, J. (1995).
Exchangeable and partially exchangeable random partitions.
Probability Theory and Related Fields, 102(2):145–158.

Pitman, J. (1996).
Some developments of the Blackwell-MacQueen urn scheme.
Lecture Notes-Monograph Series, pages 245–267.

Sethuraman, J. (1994).
A constructive definition of Dirichlet priors.
Statistica sinica, pages 639–650.

Sodjo, J., Giremus, A., Caron, F., Giovannelli, J., and Dobigeon, N. (2016).
Joint segmentation of multiple images with shared classes: A Bayesian nonparametrics approach.
In *Statistical Signal Processing Workshop (SSP), 2016 IEEE*.

Teh, Y. and Görür, D. (2009).
Indian buffet processes with power-law behavior.
In *NIPS*.

Bibliography V

Thibaux, R. and Jordan, M. (2007).
Hierarchical beta processes and the Indian buffet process.
In *International Conference on Artificial Intelligence and Statistics*, volume 11, pages 564–571.

Wood, F. and Griffiths, T. L. (2007).
Particle filtering for nonparametric Bayesian matrix factorization.
In *Advances in Neural Information Processing Systems*, volume 19, page 1513. MIT; 1998.

Xu, R., Caron, F., and Doucet, A. (2016).
Bayesian nonparametric image segmentation using a generalized Swendsen-Wang algorithm.
arXiv preprint arXiv:1602.03048.

Zhou, M., Chen, H., Ren, L., Sapiro, G., Carin, L., and Paisley, J. (2009).
Non-parametric Bayesian dictionary learning for sparse image representations.
In *Advances in neural information processing systems*, pages 2295–2303.