# Dirichlet Processes, Dependent Dirichlet Processes and Applications in Machine Learning

Zhang Zhihua

**College of Computer Science & Technology**
**Zhejiang University**
**zhzhang@zju.edu.cn**

May 2013

# Outline

# Nonparametric Bayesian Methods

- Dirichlet processes (DPs) (Ferguson, 1973; Sethuraman, 1994) or DP mixture models (Lo,1984) and Dependent DPs (MacEachern, 2000) are important nonparametric Bayesian modeling tools.

# Nonparametric Bayesian Methods

- Dirichlet processes (DPs) (Ferguson, 1973; Sethuraman, 1994) or DP mixture models (Lo,1984) and Dependent DPs (MacEachern, 2000) are important nonparametric Bayesian modeling tools.
- After Markov chain Monte Carlo (MCMC) algorithms (see, for example, Escobar and West, 1995; Bush and MacEachern, 1996; MacEachern and Muller, 1998; Neal, 2000) and Variational Bayes Algorithms (Blei and Jordan, 2005) were developed for DP mixture models, DP mixture models have been used very successfully in the literature.

# Nonparametric Bayesian Methods

- Dirichlet processes (DPs) (Ferguson, 1973; Sethuraman, 1994) or DP mixture models (Lo,1984) and Dependent DPs (MacEachern, 2000) are important nonparametric Bayesian modeling tools.
- After Markov chain Monte Carlo (MCMC) algorithms (see, for example, Escobar and West, 1995; Bush and MacEachern, 1996; MacEachern and Muller, 1998; Neal, 2000) and Variational Bayes Algorithms (Blei and Jordan, 2005) were developed for DP mixture models, DP mixture models have been used very successfully in the literature.
- Recent Developments in Statistics and Machine Learning: Duke University (David Dunson et al.), Berkeley (Mike Jordan and his students).

# Dirichlet Process Mixture Models

# Dirichlet Process Mixture Models

- In a Dirichlet Process Mixture (DPM) model, the samples $\mathbf{x}_i$ for $i = 1, \ldots, n$ are assumed to be drawn from a mixture component parameterized by $\boldsymbol{\theta}_i \in \Theta$. The $\boldsymbol{\theta}_i$s are in turn generated by the distribution $G$, which is assumed to follow a Dirichlet process prior. That is, the DPM is

$$\begin{aligned} \mathbf{x}_0 &\overset{ind}{\sim} F(\boldsymbol{\theta}_i), \quad i = 1, \ldots, n, \\ [\boldsymbol{\theta}_i | G] &\overset{iid}{\sim} G, \quad i = 1, \ldots, n, \\ G &\sim \mathrm{DP}(\alpha G_0). \end{aligned}$$

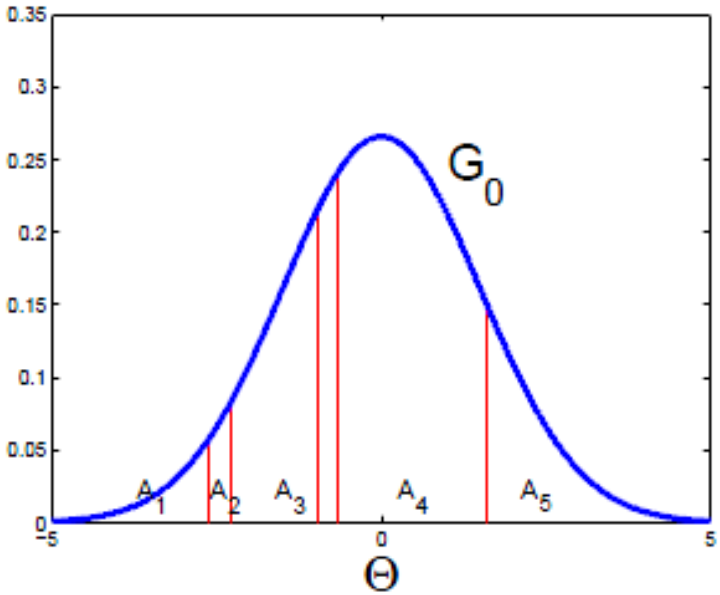
# Dirichlet Process Priors

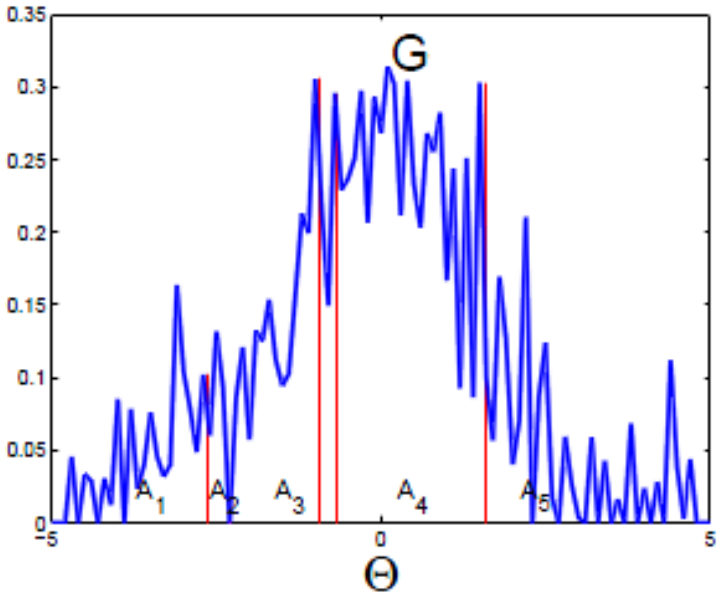
- If $G$ is drawn from the Dirichlet process $\mathrm{DP}(\alpha G_0)$ with base probability measure $G_0$ and concentration parameter $\alpha > 0$ over $(\Theta, \mathcal{A})$ then for any finite partition $(A_1, \ldots, A_k)$ of $\mathcal{A}$,

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_k)).$$

  Here $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$ denotes the Dirichlet distribution with positive parameters $\alpha_1, \ldots, \alpha_k$.

# An Illustration for DPs



# Dirichlet Processes

- Two important representations

# Dirichlet Processes

- Two important representations
  - The Pólya urn scheme (Blackwell and MacQuenn, 1973)

# Dirichlet Processes

- Two important representations
  - The Pólya urn scheme (Blackwell and MacQuenn, 1973)
  - Stick-breaking priors (Sethuraman, 1994; Pitman and Yor, 1997)

# The Pólya urn scheme

# The Pólya urn scheme

- Integrating over $G$ results in a Pólya urn scheme for the $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_1 \sim G_0(\boldsymbol{\theta}_1),$$

$$[\boldsymbol{\theta}_i | \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}] \sim \frac{\alpha G_0(\boldsymbol{\theta}_i) + \sum_{l=1}^{i-1} \delta(\boldsymbol{\theta}_i | \boldsymbol{\theta}_l)}{\alpha + i - 1},$$

where $\delta(\boldsymbol{\theta}_i | \boldsymbol{\theta}_l)$ is a point mass at $\boldsymbol{\theta}_l$.

# The Pólya urn scheme

- Integrating over $G$ results in a Pólya urn scheme for the $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_1 \sim G_0(\boldsymbol{\theta}_1),$$

$$[\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}] \sim \frac{\alpha G_0(\boldsymbol{\theta}_i) + \sum_{l=1}^{i-1} \delta(\boldsymbol{\theta}_i|\boldsymbol{\theta}_l)}{\alpha + i - 1},$$

  where $\delta(\boldsymbol{\theta}_i|\boldsymbol{\theta}_l)$ is a point mass at $\boldsymbol{\theta}_l$.
- We see that as $\alpha \to 0$, all the $\boldsymbol{\theta}_i$ are identical to $\boldsymbol{\theta}_1$, which in turn follows $G_0$. When $\alpha \to \infty$, the $\boldsymbol{\theta}_i$ become iid $G_0$.

# The Pólya urn scheme

- Integrating over $G$ results in a Pólya urn scheme for the $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_1 \sim G_0(\boldsymbol{\theta}_1),$$

$$[\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}] \sim \frac{\alpha G_0(\boldsymbol{\theta}_i) + \sum_{l=1}^{i-1} \delta(\boldsymbol{\theta}_i|\boldsymbol{\theta}_l)}{\alpha + i - 1},$$

  where $\delta(\boldsymbol{\theta}_i|\boldsymbol{\theta}_l)$ is a point mass at $\boldsymbol{\theta}_l$.
- We see that as $\alpha \to 0$, all the $\boldsymbol{\theta}_i$ are identical to $\boldsymbol{\theta}_1$, which in turn follows $G_0$. When $\alpha \to \infty$, the $\boldsymbol{\theta}_i$ become iid $G_0$.
- Since the $\boldsymbol{\theta}_i$ are exchangeable, the Pólya urn scheme can be written as

$$[\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}] \sim \frac{\alpha G_0(\boldsymbol{\theta}_i) + \sum_{l \neq i} \delta(\boldsymbol{\theta}_i|\boldsymbol{\theta}_l)}{\alpha + n - 1},$$

  were $\boldsymbol{\theta}_{-i}$ represents $\{\boldsymbol{\theta}_l : l \neq i\}$.

# The clustering property of the DP prior

- The discreteness of the random distribution $G$ leads to the well-known clustering property of the DP, which plays a central role in nonparametric Bayesian inference and computation.

# The clustering property of the DP prior

- The discreteness of the random distribution $G$ leads to the well-known clustering property of the DP, which plays a central role in nonparametric Bayesian inference and computation.
- The clustering property allows DPs to formalize the notion of "borrowing strength" across related studies.

# The clustering property of the DP prior

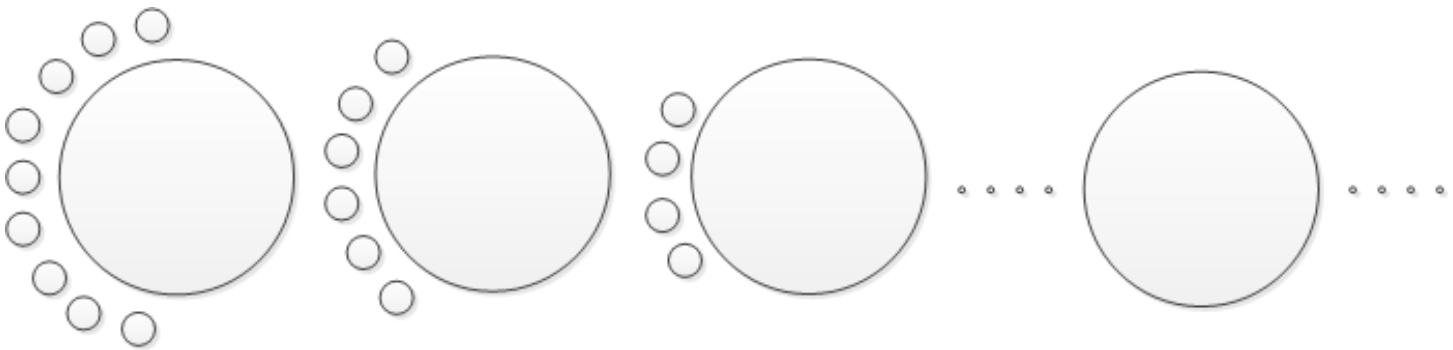
- Assume that there are $c$ distinct values among the $\boldsymbol{\theta}_i$ as $\Phi = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_c\}$, and that there are $n_k$ occurrences of $\boldsymbol{\phi}_k$ such that $\sum_{k=1}^{c} n_k = n$. The vector of configuration indicators, $\mathbf{w} = (w_1, \ldots, w_n)$, is defined by $w_i = k$ if and only if $\boldsymbol{\theta}_i = \boldsymbol{\phi}_k$ for $i = 1, \ldots, n$. Thus $(\Phi, \mathbf{w})$ is an equivalent representation of $\boldsymbol{\Theta}$, and hence the DP is also defined as

$$[\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}] \sim \frac{\alpha G_0(\cdot) + \sum_{k=1}^{c} n_{k(-i)} \delta(\boldsymbol{\theta}_i | \boldsymbol{\phi}_k)}{\alpha + n - 1},$$

where $n_{k(-i)}$ refers to the cardinality of cluster $k$, with $\boldsymbol{\theta}_i$ removed, and

$$\boldsymbol{\phi}_k \overset{iid}{\sim} G_0(\cdot), \; k = 1, \ldots, c.$$

# An Alternative View: Chinese Restaurant Processes



# Stick-Breaking Priors

- The Stick-Breaking Prior is defined by

$$P(\cdot) = \sum_{k=1}^{K} w_k \delta(\cdot|\phi_k), \quad \phi_k \overset{iid}{\sim} G_0$$

$$w_1 = q_1 \text{ and } w_k = q_k \prod_{l=1}^{k-1}(1 - q_l), \quad q_k \sim \text{Beta}(a_k, b_k)$$

where the number of atoms $K$ can be finite or infinite.

# Stick-Breaking Priors

- The Stick-Breaking Prior is defined by

$$P(\cdot) = \sum_{k=1}^{K} w_k \delta(\cdot|\phi_k), \quad \phi_k \overset{iid}{\sim} G_0$$

$$w_1 = q_1 \text{ and } w_k = q_k \prod_{l=1}^{k-1}(1 - q_l), \quad q_k \sim \text{Beta}(a_k, b_k)$$

  where the number of atoms $K$ can be finite or infinite.

- In the finite case, setting $q_K = 1$ guarantees that $\sum_{k=1}^{K} w_k = 1$ with probability 1.

# Stick-Breaking Priors

- In the infinite case,

$$\sum_{k=1}^{\infty} w_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} \mathbb{E}(\log(1 - q_k)) = -\infty.$$

Alternatively, it is sufficient to check that $\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty$ (Ishwaran and James, 2001).

# Stick-Breaking Priors

- In the infinite case,

$$\sum_{k=1}^{\infty} w_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} \mathbb{E}(\log(1 - q_k)) = -\infty.$$

  Alternatively, it is sufficient to check that $\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty$ (Ishwaran and James, 2001).
- For example, $a_k = 1 - a$ and $b_k = b + ka$ for $0 \leq a \leq 1$ and $b > -a$ leads to the two-parameter Poisson-Dirichlet process, as known as the Pitman-Yor process.

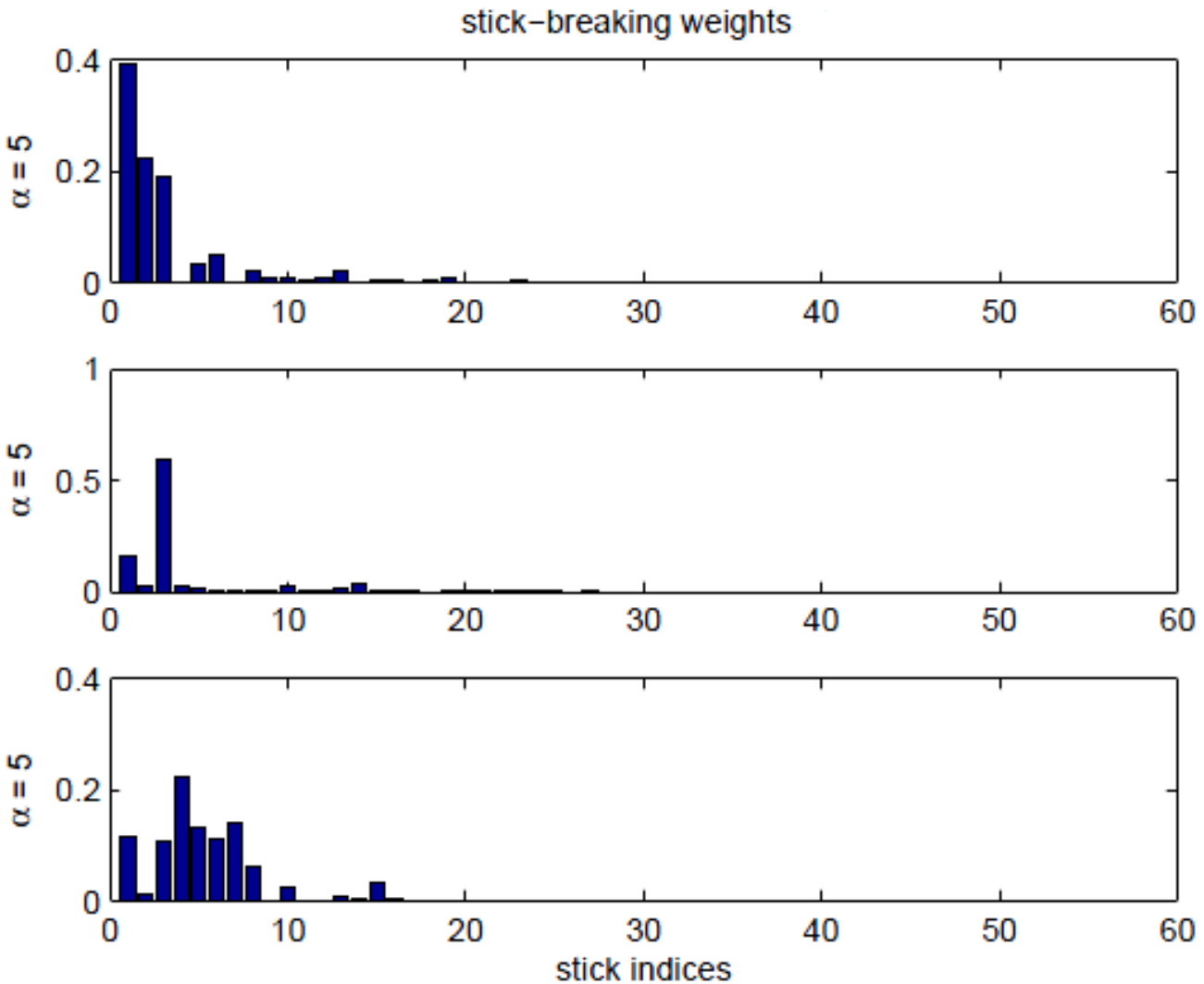
# Stick-Breaking Priors

- In the infinite case,

$$\sum_{k=1}^{\infty} w_k = 1 \text{ a.s. iff } \sum_{k=1}^{\infty} \mathbb{E}(\log(1-q_k)) = -\infty.$$

  Alternatively, it is sufficient to check that $\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty$ (Ishwaran and James, 2001).
- For example, $a_k = 1 - a$ and $b_k = b + ka$ for $0 \leq a \leq 1$ and $b > -a$ leads to the two-parameter Poisson-Dirichlet process, as known as the Pitman-Yor process.
- Especial cases: $q_k \overset{iid}{\sim} \mathrm{Beta}(1, \alpha)$ where $\alpha > 0$, that is, $a = 0$ and $b = \alpha$ (Sethuraman, 1994); $q_k \overset{iid}{\sim} \mathrm{Beta}(1-a, ka)$ (Pitman and Yor, 1997).

# An Illustration for Stick-Breaking Priors



# The Stick-Breaking Representation and Pólya urn scheme

- The Pólya urn scheme makes inference and computation more feasible. Moreover, clustering property of the DP is very useful in practical applications. For example, it can be used for automatical choice of the number of classes in clustering analysis.

# The Stick-Breaking Representation and Pólya urn scheme

- The Pólya urn scheme makes inference and computation more feasible. Moreover, clustering property of the DP is very useful in practical applications. For example, it can be used for automatical choice of the number of classes in clustering analysis.
- The Stick-Breaking Representation makes modeling more powerful. But the corresponding computation usually requires a truncated technique in the infinite case.

# Dependent Dirichlet Processes

# Dependent Dirichlet Processes

- Dependent DPs (DDPs) provide a general framework to describe dependency among a collection of stochastic processes.

# Dependent Dirichlet Processes

- Dependent DPs (DDPs) provide a general framework to describe dependency among a collection of stochastic processes.
- A principled approach to this direction is to treat the weights in the stick-breaking representation as stochastic functions (Delorio et al, 2004).

# Dependent Dirichlet Processes

# Dependent Dirichlet Processes

- Other ways of achieving dependence among random measures include the hierarchical DP model (Teh et al, 2006), the use of linear combinations of realizations of independent DPs (Muller et al, 2004) and kernel-weighted mixture of DPs (Dunson et al, 2007). These are specialized approaches that can make use of generalized Pólya urn schemes for posterior inference and prediction.

# Dependent Dirichlet Processes

- Other ways of achieving dependence among random measures include the hierarchical DP model (Teh et al, 2006), the use of linear combinations of realizations of independent DPs (Muller et al, 2004) and kernel-weighted mixture of DPs (Dunson et al, 2007). These are specialized approaches that can make use of generalized Pólya urn schemes for posterior inference and prediction.
- It is also desirable to address this issue under the "single-$p$" DPP (MacEachern, 2000), because inference and computation for the resulting model can proceed via a relatively straightforward application of the well-established MCMC techniques.

# Dependent Dirichlet Processes

# Dependent Dirichlet Processes

- Assume there a collection of a collection of random probability measures $G_j$ on $(\Phi, \mathcal{B})$.
- In general, there are two extreme constructions for the $G_j$. For the first construction, $G_j$ are treated as independent DPs given hyperparameters $\theta$, so the model is equivalent to the $m$ separate submodels. For the second one, the model is treated as a single conventional DP, i.e., $G_1 = \cdots = G_q$. Clearly, the first case allows too little sharing of strength in many applications, while the second case enforces too much sharing.

# The Kernel Weighted Mixture of DPs

- Assume there a set of samples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ where $\mathbf{x}_i \in \mathbb{R}^p$ is an input vector and $y_i$ is the corresponding output.
- The kernel weighted mixture of DPs (Dunson et al, 2007) is specified as

$$G_{\mathbf{x}} = \sum_{l=1}^{n} b_l(\mathbf{x}) G_l^*, \; G_l^* \overset{iid}{\sim} \mathrm{DP}(\alpha G_0), \text{ for } l = 1, \ldots, n,$$

where $b_l(\mathbf{x})$ is a kernel-based weight.

# Capture relationships among multiple studies

- In order to model relationships among multiple studies, we consider a nonparametric hierarchical model. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})^T$, $\mathbf{y}_{\cdot j} = (y_{1j}, \ldots, y_{n_j j})^T$ denote the response vector in study $j$. The model is

$$[y_{ij}|\mathbf{b}_{ij}] \overset{ind}{\sim} F(y_{ij}|\mathbf{b}_{ij}),\ j = 1, \ldots, q \text{ and } i = 1, \ldots, n_j;$$
$$[\mathbf{b}_{ij}|G_j] \overset{iid}{\sim} G_j,\ i = 1, \ldots, n_j \text{ for each } j.$$

# The Nested Dirichlet Process

- The Nested Dirichlet Process (Rodriguez et al., 2008) is defined by

$$\begin{aligned} G_j(\cdot) &\sim Q \triangleq \sum_{k=1}^{\infty} \pi_k^* \delta(\cdot|G_k^*), \\ G_k^*(\cdot) &\triangleq \sum_{l=1}^{\infty} w_{lk} \delta(\cdot|\phi_{lk}), \\ \phi_{lk} &\sim G_0, \end{aligned}$$

with $w_{lk} = u_{lk} \prod_{s=1}^{l-1}(1-u_{sk})$, $\pi_k^* = v_k \prod_{s=1}^{k-1}(1-v_s)$, $v_k \sim \mathrm{Beta}(1,\alpha)$, and $u_{lk} \sim \mathrm{Beta}(1,\beta)$.

# The Conditional Autoregressive DP

- We model the $G_j$ in autoregressive form of

$$G_j = \omega_{jj} G_j^* + \sum_{l \neq j} \omega_{jl} G_l, \quad j = 1, \ldots, q, \tag{1}$$

where $0 \leq \omega_{jl} < 1$ and $\sum_{l=1}^{q} \omega_{jl} = 1$.

- From (1), we get the following conditional autoregressive model

$$E(G_j(A)|G_l(A), l \neq j) = \omega_{jj} G_0(A) + \sum_{l \neq j} \omega_{jl} G_l(A)$$

for any Borel set $A \in \mathcal{A}$. We thus say the $G_j$ defined by (1) follow *conditional autoregressive* DPs.

# Graphical Representations

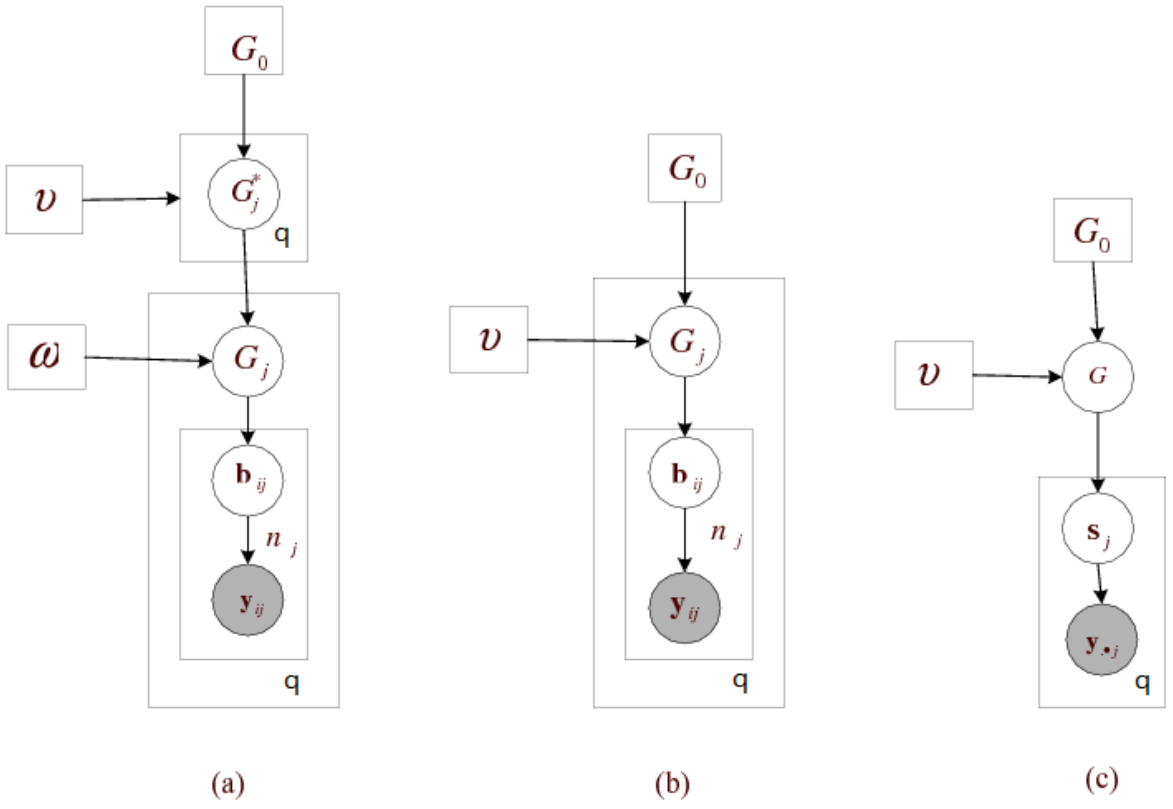



Figure: Graphical Representations: (a) the conditional autoregressive DP model, (b) $q$ independent DP mixture models, and (c) DP mixture model.

# The Spatial DP Model

- The spatial DP (sDP) model (Gelfand et al., 2005) is

$$\begin{aligned} [\mathbf{y}_{\cdot j}|\mathbf{s}_j] &\overset{ind}{\sim} F(\cdot|\mathbf{s}_j), \quad j = 1, \ldots, q, \\ [\mathbf{s}_j|G] &\overset{iid}{\sim} G, \quad j = 1, \ldots, q, \\ [G|\alpha, G_0] &\sim \mathrm{DP}(\alpha G_0). \end{aligned}$$

Furthermore, the base distribution is defined as a Gaussian process (GP). Specifically, this model describes the dependence among the response variates via DP, and the dependence among the instances via GP.

# The Matrix-Variate DP Model

- The Matrix-Variate DP mixture (Zhang et al., 2010) is

$$[\mathbf{y}_i|\mathbf{\Theta}_i] \overset{ind}{\sim} F(\mathbf{\Theta}_i), \quad i = 1, \ldots, n,$$
$$[\mathbf{\Theta}_i|G] \overset{iid}{\sim} G, \quad i = 1, \ldots, n,$$
$$G \sim \mathrm{DP}(\alpha G_0).$$

  Here $\mathbf{\Theta}_i$'s are a collection of matrices of the same size (e.g., $p \times q$), and we assume that the base probability measure $G_0$ follows a matrix-variate distribution. We thus refer to the resulting DP as a *matrix-variate DP* (MATDP).

- As a concrete example, let $G_0$ follow a matrix-variate normal distribution of the form

$$G_0(\cdot|\mathbf{\Sigma}, \mathbf{\Lambda}) = N_{p,q}(\cdot|\mathbf{M}, \ \mathbf{A} \otimes \mathbf{B}).$$

# Graphical Representations

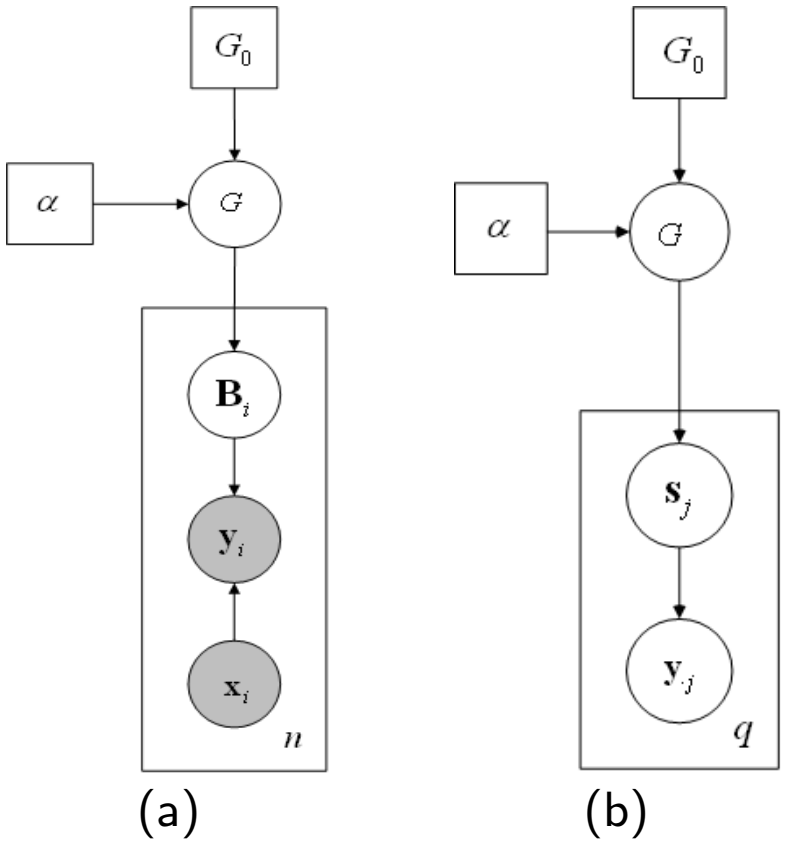
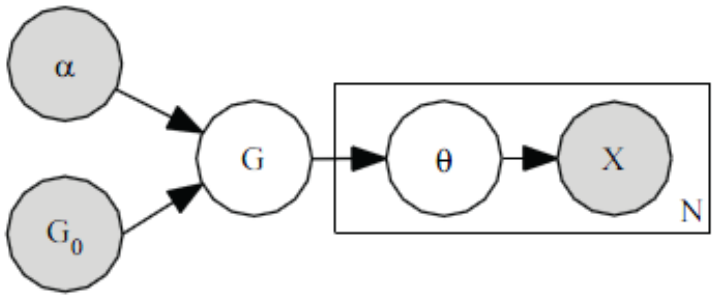



Figure: Graphical representations under regression setting: (a) MATDP and (b) sDP.

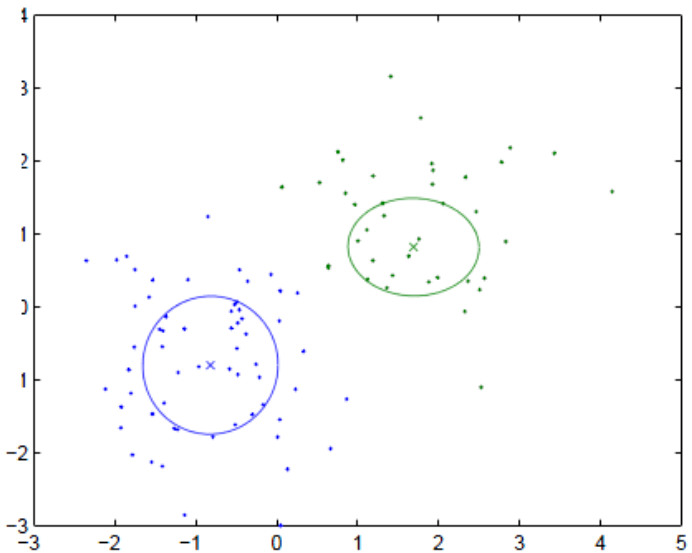
# Dirichlet Process Mixture of Gaussian Density

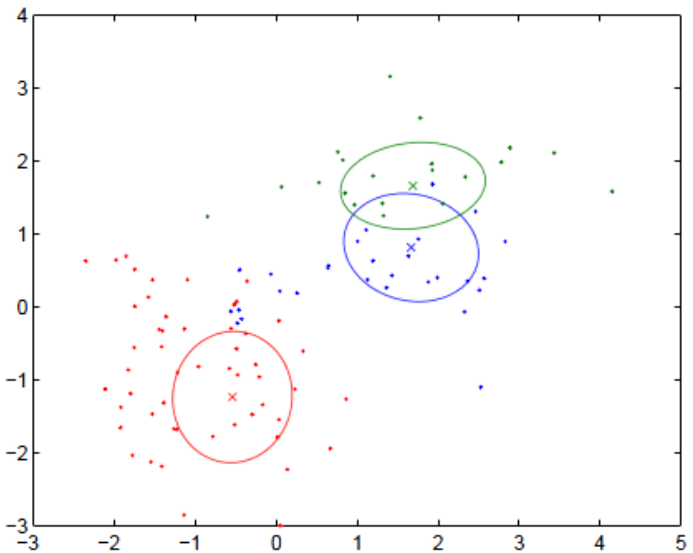



- Let $\boldsymbol{\theta}_i = (\mu_i, \Sigma_i)$ be the parameter of Gaussians for $i = 1, \ldots, N$;
- Let $\mathbf{x}_i \sim N(\cdot|\boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i \sim G$ and $G \sim \mathrm{DP}(\alpha G_0)$;
- $G_0$ is defined as a Normal-Inverse Wishart *NIW*.

# Dirichlet Process Mixture of Gaussian Density

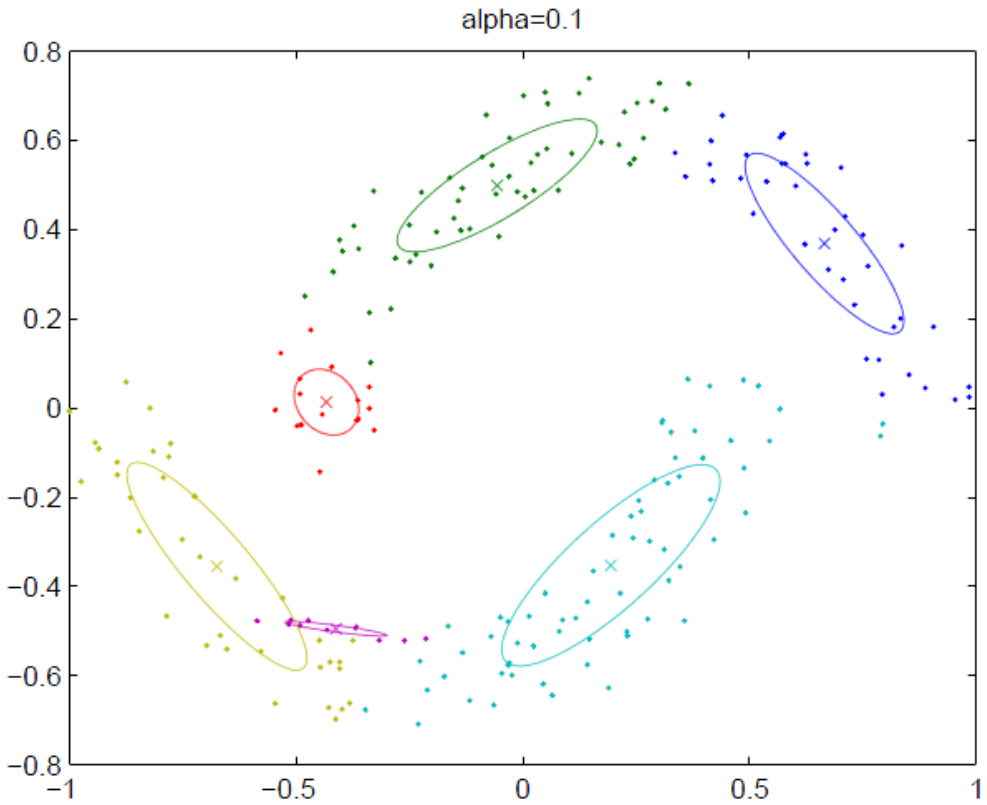
- Different values of $\alpha$ yield different number of components.



# Dirichlet Process Mixture of Gaussian Density

- The Two Moon Data: a more illustrative example for automatically determining the number of components.



# Nonlinear Supervised Learning Models

- We consider a regression problem on a training dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_1^n$, where $\mathbf{x}_i \subset \mathbb{R}^p$ is an input or covariate vector and $\mathbf{y}_i \in \mathbb{R}^q$ is a $q$-dimensional continuous vector of responses, and we also treat multivariate classification problems, specifically the multi-class classification problem and the multi-label prediction problem. In the latter problem, the label associated with an input vector $\mathbf{x}$ is $\mathbf{y} \in \{0, 1\}^q$. Unlike the conventional multi-class classification problem in which $\mathbf{x}$ belongs to one and only one class, in the multi-label problem $\mathbf{x}$ is allowed to belong simultaneously to several classes.

# Nonlinear Supervised Learning Models

- We are currently interested in jointly modeling two types of relationships: the dependency among the data instances and the dependency among the response (or input) variates. This is a challenging and interesting issue, because it provides leverage on problems where the data are not iid (independent and identically distributed) while the (co)variates are not independent.

# Dirichlet Process Multinomial Logit (dpMNL) models

The specification of dpMNL (Shahbaba and Neal, 2009) is

$$\begin{aligned}
(\mathbf{x}_i, y_{ij})|\mathbf{b}_{ij}, \sigma^2, \mu_i, \Sigma_i &\overset{ind}{\sim} MNL(y_{ij}|\mathbf{x}_i^T \mathbf{b}_{ij}) \times \\
&\quad N_n(\mathbf{x}_i|\mu_i, \Sigma_i), i = 1, \ldots, n; \\
(\mathbf{b}_{ij}, \mu_i, \Sigma_i)|G_j &\overset{iid}{\sim} G_j, \quad i = 1, \ldots, n; \\
G_j|\nu, G_0 &\overset{iid}{\sim} \mathrm{DP}(\alpha G_0).
\end{aligned}$$

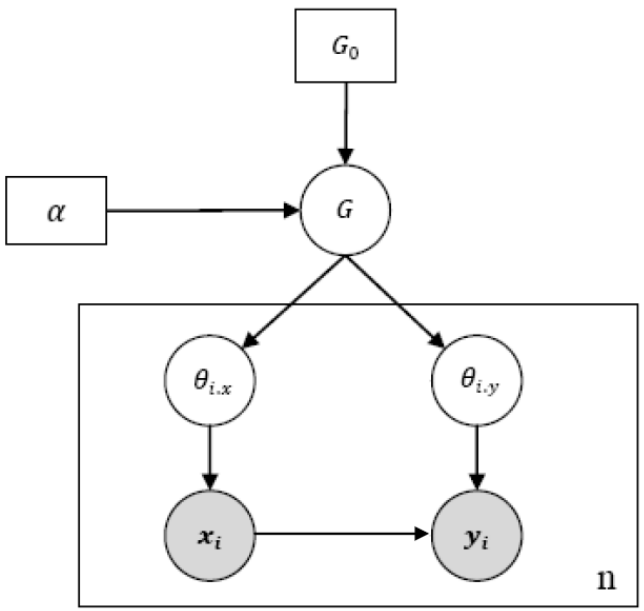
# Dirichlet Process Latent Factor Models (DP-LFM)

The specification of DP-LFM (Zhang et al., 2013) is

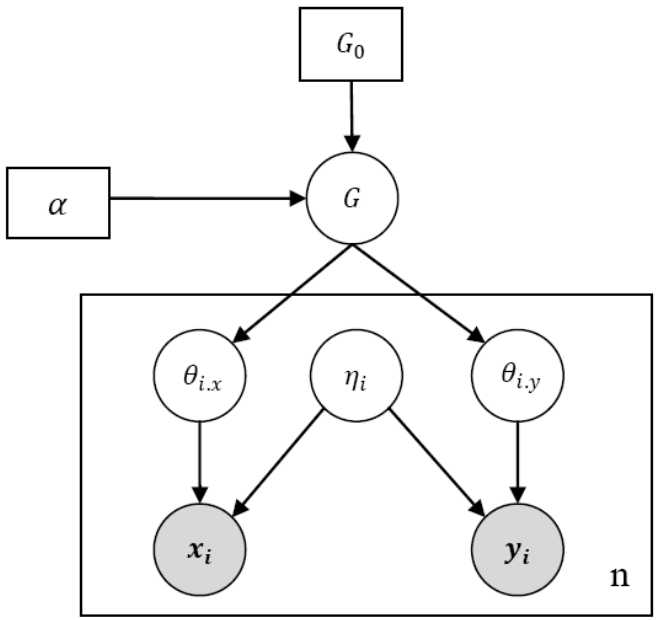
$$\begin{aligned}
\mathbf{x}_i &\sim N(\cdot|\mathbf{A}_i\boldsymbol{\eta}_i + \boldsymbol{\mu}_i,\ \boldsymbol{\Sigma}_i),\\
\mathbf{y}_i &\sim F(\cdot|\mathbf{B}_i\boldsymbol{\eta}_i + \boldsymbol{\nu}_i,\ \boldsymbol{\Lambda}_i),\\
\boldsymbol{\eta}_i &\sim N(\cdot|\mathbf{0}, \mathbf{I}_r)\\
\boldsymbol{\theta}_i|G &\sim G(\cdot),\\
G &\sim DP(\alpha G_0),
\end{aligned}$$

where $\boldsymbol{\Sigma}_i = \mathrm{diag}(\sigma_{i1}^2, \ldots \sigma_{ip}^2)$, $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}^2, \ldots, \lambda_{iq}^2)$, and the $\boldsymbol{\theta}_i = \{\mathbf{A}_i, \mathbf{B}_i, \boldsymbol{\mu}_i, \boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\Lambda}_i\}$ are the parameters following a joint distribution generated from the DP prior $\mathrm{DP}(\alpha G_0)$.

# Dirichlet Process Mixture of Gaussian Density




(a) dpMNL




(b) DP-LFM

# Summary

- Introduce two definitions of DP priors: Pólyn urn scheme and stick-braking representation.
- Introduce modeling approaches to DDPs.
- Some Applications in Machine Learning.

# Several Active Issues

- Extensions: Beta Processes (Indian Buffet Processes) and Pólyn Tree Processes

# Several Active Issues

- Extensions: Beta Processes (Indian Buffet Processes) and Pólyn Tree Processes
- Efficient Computation: Sequential Monte Carlo and Online Variational EM algorithms.

# References: DP Models

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

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

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

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

5 A. Y. Lo. On a class of Bayesian nonparametric estimates: I. density estimates. The Annals of Statistics 12 (1), 351–357, 1984.

# References: Computations

6 C. A. Bush and S. N. MacEachern. A semiparametric Bayesian model for randomised block designs. Biometrika, 83:275–285, 1996.

7 M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90:577–588, 1995.

8 S. N. MacEachern. Computational methods for mixture of Dirichlet process models. In D. Dey, P.Muller and D. Sinha, editors, Practical Nonparametric and Semiparametric Bayesian Statistics, pages 23–43. Springer-Verlag, New York, 1998.

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

# References: Computations

10 S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics, 13: 158–182, 2004.

11 S. Jain and R. M. Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. Bayesian Analysis, 2(3): 445–472, 2007.

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

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

14 D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1: 121–144, 2005.

# References: DDPs

15 S. N. MacEachern. Dependent nonparametric processes. In The Section on Bayesian Statistical Science, pages 50–55. American Statistical Association, 1999.

16 M. De Iorio, P. Muller, G. L. Rosner and S. N. MacEachern. An ANOVA model for dependent random measures. Journal of the American Statistical Association, 99:205–215, 2004.

17 J. E. Griffin and M. F. J. Steel. Order-based dependent Dirichlet processes. Journal of the American Statistical Association, 101(473):179–194, 2006.

18 P. Muller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. Journal of the Royal Statistical Society Series B, 66(3):735–749, 2004.

# References: DDPs

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

20 D. B. Dunson, N. Pillai, and J.-H. Park. Bayesian density regression. Journal of the Royal Statistical Society Series B, 69(2):163–183, 2007.

21 A. Rodriguez, D. B. Dunson and A. E. Gelfand. The Nested Dirichlet processes (with discussion). Journal of the American Statistical Association, 103:1131–1154, 2008.

22 Zhihua Zhang, D. Wang and E. Y. Chang. An Autoregressive Approach to Nonparametric Hierarchical Dependent Modeling. The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP 11, 2012.

# References: DDPs

23 A. E. Gelfand, A. Kottas and S. N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. Journal of the American Statistical Association, 100:1021–1035, 2005.

24 Zhihua Zhang, G. Dai and M. I. Jordan. Matrix-Variate Dirichlet Process Mixture Models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP 9, 2010.

# References: Applications

25 E. Sudderth and M. I. Jordan. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In Advances in Neural Information Processing Systems, volume 21, Cambridge, MA. MIT Press, 2009.

26 B. Shahbaba and R. Neal. Nonlinear models using Dirichlet process mixtures. Journal of Machine Learning Research. 10(2): 1829–1850, 2009.

27 P. D. Hoff. Model-based subspace clustering. Bayesian Analysis. 1(2): 321–344, 2006.

28 Y. Xue, X. Liao and L. Carin. Multi-task learning for classification with Dirichlet process priors. Journal of Machine Learning Research, 8: 35–63, 2007.

29 Zhihua Zhang, Dakan Wang, Guang Dai and Michael I. Jordan. Matrix-Variate Dirichlet Process Mixture Models with Applications. Technical Report, 2013.

# Thanks

Questions & Comments!