Advanced Bayesian Learning
Bayesian Nonparametrics
Spring 2014

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Topic overview

- Reminder: Multinomial data - Dirichlet prior
- Bayesian histograms
- Dirichlet process
- Beyond the Dirichlet process: Pitman-Yor and Probit stick-breaking
- The Dirichlet process mixtures
- MCMC for Dirichlet process mixtures

The Dirichlet distribution

- $\theta \sim \text{Dirichlet}(a_1, ..., a_k)$ with density

$$p(\theta_1, \theta_2, ..., \theta_k) \propto \prod_{j=1}^{k} \theta_j^{a_j - 1}.$$

- Define $\alpha = \sum_{j=1}^{k} a_j$ and $\pi_0 = a/\alpha$.
- **Expected value** and **variance** of the *Dirichlet*$(a_1, ..., a_k)$ distribution

$$\mathrm{E}(\theta_j) = \frac{a_j}{\alpha} = \pi_{0j} \qquad \mathrm{V}(\theta_j) = \frac{\mathrm{E}(\theta_j)\left[1 - \mathrm{E}(\theta_j)\right]}{1 + \alpha}$$

- Note that α is a **precision** parameter (large α means low variance).

Conjugate analysis for multinomial data

- **Data**: $y = (n_1, ..., n_k)$, where n_j = number of items in category j.
- **Prior**
$$\theta \sim \text{Dirichlet}(a_1, ..., a_k)$$
- **Likelihood**
$$p(n_1, n_2, ..., n_k | \theta_1, \theta_2, ..., \theta_k) \propto \prod_{j=1}^{k} \theta_j^{n_j}$$
- **Posterior**
$$\theta | n_1, ..., n_k \sim \text{Dirichlet}(n_1 + a_1, ..., n_k + a_k)$$
- **Posterior expected value**
$$E(\theta_j | n_1, ..., n_k) = \frac{n_j + a_j}{n + \alpha}$$

Bayesian histograms

- Histogram partitions the data space $\xi_0 < \xi_1 < ... < \xi_k$ and records how many observations end up in each bin (B_h). Multinomial data.
- Probability model for **histograms**

$$f(y) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_h} \left(\frac{\pi_h}{\xi_h - \xi_{h-1}} \right)$$

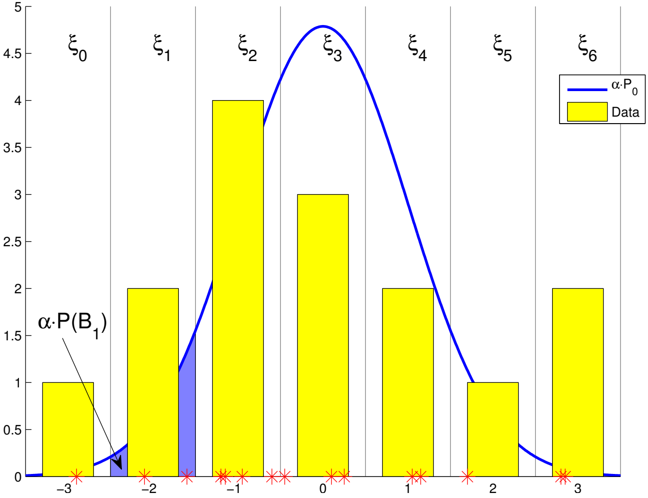
- n_h = number of data points in partition (bin) h: $\xi_{h-1} < y \leq \xi_h$.
- **Prior** on $\pi = (\pi_1, ..., \pi_k)$

$$\pi \sim Dirichlet(a_1, ..., a_k)$$

- **Posterior**

$$\pi | n_1, ..., n_k \sim \text{Dirichlet}(n_1 + a_1, ..., n_k + a_k)$$

Illustration of Bayesian histograms



Bayesian histograms, cont.

- **Posterior**

$$\pi | n_1, ..., n_k \sim \text{Dirichlet}(n_1 + a_1, ..., n_k + a_k)$$

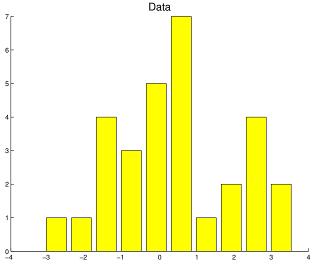
- Specify $a_1, .., a_k$ through $\pi_0 = (\pi_{01}, ..., \pi_{0k})$ and $\alpha = \sum_{j=1}^{k} a_j$.
- Specify π_0 from a **base distribution** P_0. For the hth bin:

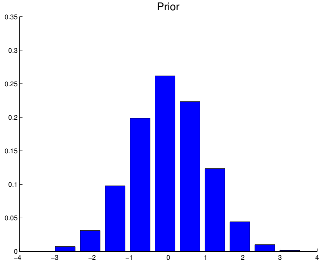
$$\pi_{0h} = P_0(B_h) = \Pr(\xi_{h-1} < y \le \xi_h)$$

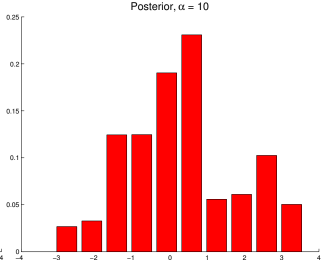
.

- The Dirichlet prior is a **computational dream**, and it is **easy to specify** the hyperparameters π_0 and α.
- But, the Dirichlet prior **lacks smoothness**: all pairs of bins have negative correlations, regardless of how near they are.
- Sensitive to the choice of bins.

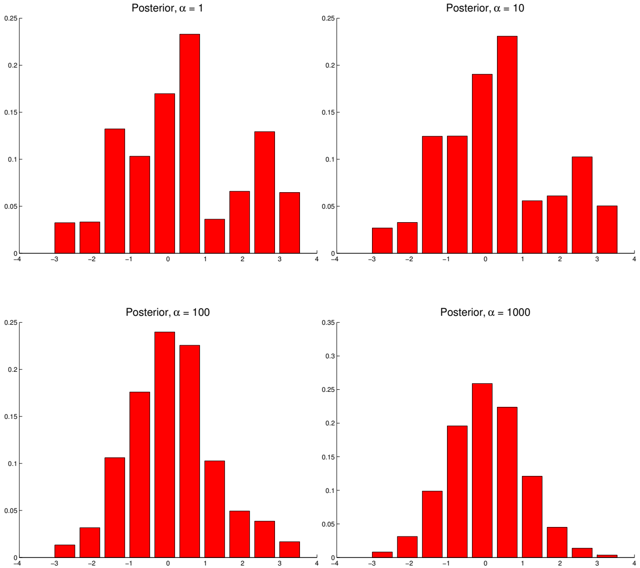
Bayesian histogram example



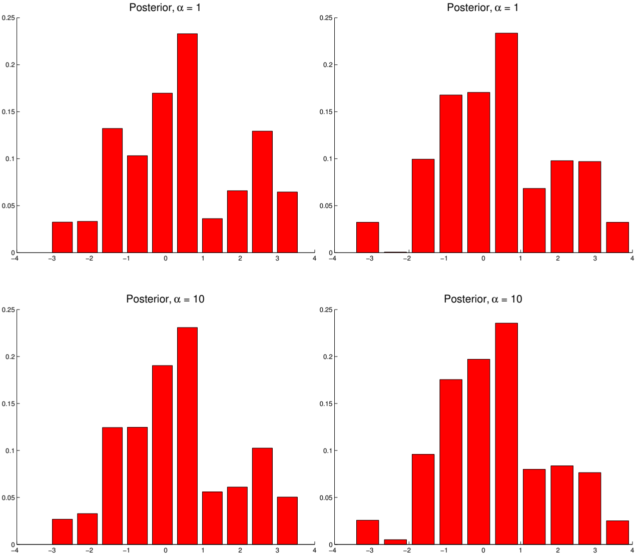




Bayesian histogram example



Histograms are sensitive to the choice of bins



The Dirichlet process

- Let $B_1, B_2, ..., B_k$ be a partition of the outcome space Ω.
- Let $P(B_1), ..., P(B_k)$ denote the distribution over the partition.
- Dirichlet distribution is a **distribution over a space of distributions**:

$$(P(B_1), ..., P(B_k)) \sim Dirichlet(\alpha P_0(B_1), ..., \alpha P_0(B_k))$$

  where P_0 is a fixed probability measure (e.g. the $N(0, 1)$ density).
- Dirichlet distribution is closed under summation or splitting of bins.
- Can be used to define a **stochastic process** in a consistent way. Compare with GPs.
- A random probability measure P follows a **Dirichlet process** $P \sim DP(\alpha \cdot P_0)$ with base measure P_0 and precision parameter α iff

$$(P(B_1), ..., P(B_k)) \sim Dirichlet(\alpha P_0(B_1), ..., \alpha P_0(B_k))$$

  for any finite (measureable) partition $B_1, ..., B_k$.

The Dirichlet process - properties

- If $P \sim DP(\alpha P_0)$ then

$$P(B) \sim \mathrm{Beta}\left[\alpha P_0(B), \alpha\left(1 - P_0(B)\right)\right], \text{ for any } B \in \mathcal{B}$$

$$E\left[P(B)\right] = P_0(B)$$

$$\mathrm{Var}\left[P(B)\right] = P_0(B)\left[1 - P_0(B)\right]/(1+\alpha)$$

- **Model**

$$y_i | P \overset{iid}{\sim} P \text{ , for } i = 1, ..., n$$

- **Prior**

$$P \sim DP(\alpha P_0)$$

- **Posterior** for a **finite partition**, $P(B_1), ..., P(B_k)|\mathbf{y}$ is

$$Dir\left(\alpha P_0(B_1) + \sum_{i=1}^{n} 1_{y_i \in B_1}, ..., \alpha P_0(B_k) + \sum_{i=1}^{n} 1_{y_i \in B_k}\right)$$

THE DIRICHLET PROCESS - PROPERTIES

- **Posterior** for the unknown probability distribution P

$$P|y_1, ..., y_n \sim DP\left(\alpha P_0 + \sum_{i=1}^{n} \delta_{y_i}\right)$$

- Since

$$P(B) \sim Beta\left(\alpha P_0(B) + \sum_{i=1}^{n} 1_{y_i \in B}, \alpha(1 - P_0(B)) + \sum_{i=1}^{n} 1_{y_i \in B^c}\right)$$

so

$$E\left(P(B)|y_1, ..., y_n\right) = \left(\frac{\alpha}{\alpha + n}\right) P_0(B) + \left(\frac{n}{\alpha + n}\right) \sum_{i=1}^{n} \frac{1}{n} 1_{y_i \in B^c}$$

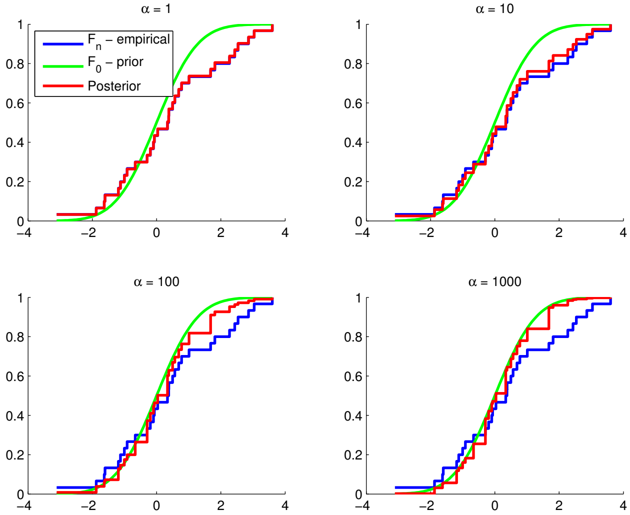
Estimating a d.f. with a DP prior

- If $B = (-\infty, y]$ then

$$E\left(F(y) | y_1, ..., y_n\right) = \left(\frac{\alpha}{\alpha + n}\right) F_0(y) + \left(\frac{n}{\alpha + n}\right) F_n(y)$$

- where
  - $F(y)$ is the unknown d.f.
  - $F_0(y)$ is the d.f. from P_0
  - $F_n(y) = \frac{1}{n} \sum 1_{y_i \leq y}$ is the empirical d.f.
- Note: under the DP posterior, $F(\cdot)$ is discrete with probability one. Not great for continuous data ...
- This is true in general: **realisations from a DP are discrete with probability one**.

Estimating a d.f. with a DP prior



Stick-breaking characterization of the DP

- $P \sim DP(\alpha P_0)$ is equivalent to an **infinite mixture of point masses**

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_i}$$

$$\pi_h = V_h \prod_{\ell < h} (1 - V_\ell)$$

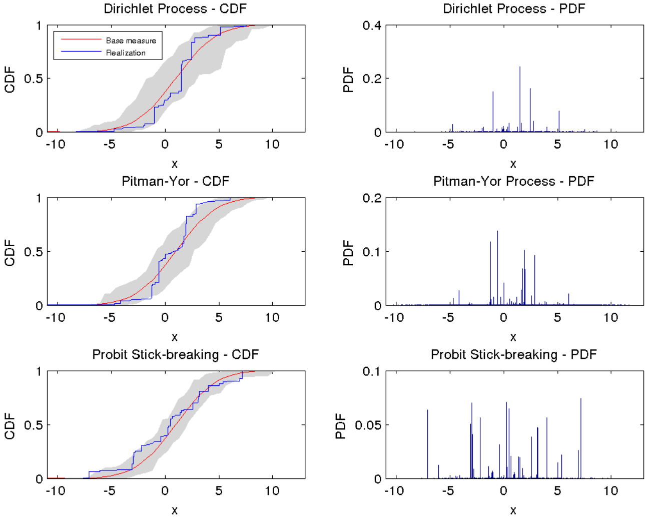
$$V_h \overset{iid}{\sim} Beta(1, \alpha)$$

$$\theta_h \overset{iid}{\sim} P_0$$

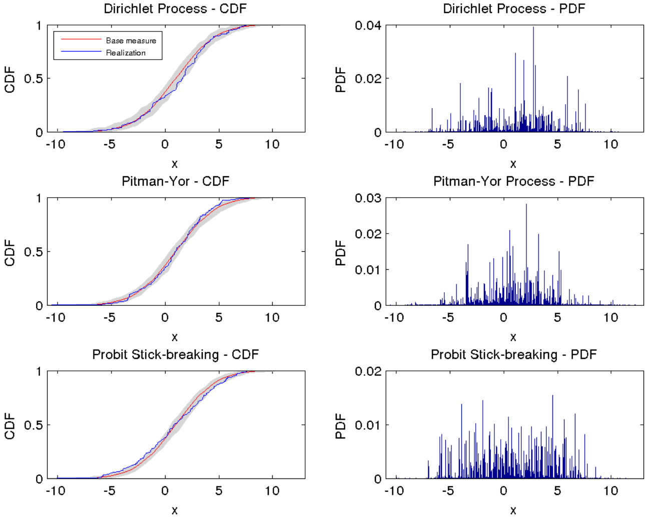
- Stick picture
- Alternative notation for $P \sim DP(\alpha P_0)$:

$$\pi = (\pi_1, \pi_2, ...) \sim \text{Stick}(\alpha) \text{ and } \theta_h \sim P_0$$

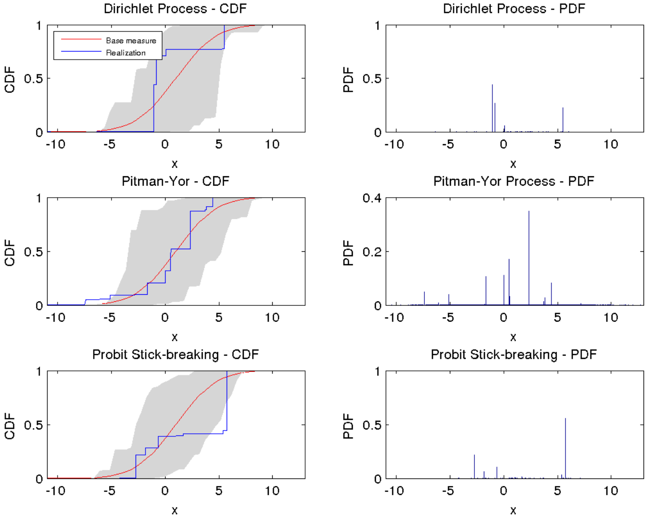
SIMULATING STICK-BREAKING PRIORS $\alpha = 10$



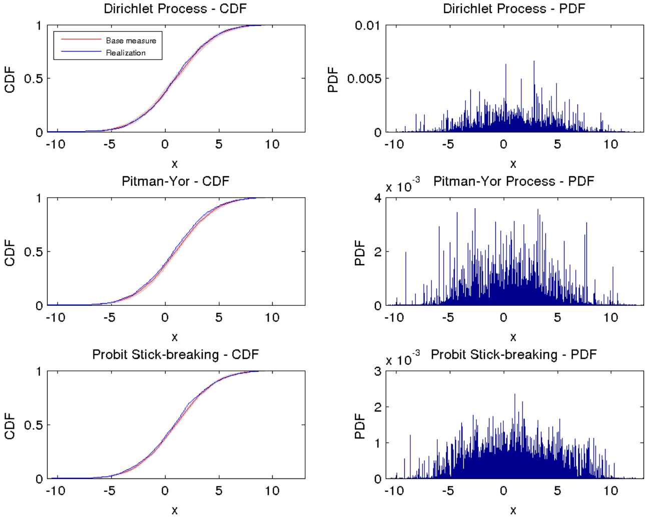
SIMULATING STICK-BREAKING PRIORS $\alpha = 100$



SIMULATING STICK-BREAKING PRIORS $\alpha = 1$



Simulating stick-breaking priors $\alpha = 1000$



Beyond DP - Pitman-Yor and Probit sticks

- **Pitman-Yor process** with parameters P_0, $0 \leq a < 1$ and $b > -a$:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_i} \quad \theta_h \overset{iid}{\sim} P_0$$

$$\pi_h = V_h \prod_{\ell<h} (1 - V_\ell)$$

$$V_h \overset{iid}{\sim} Beta(1 - a, b + ha)$$

- **Probit stick-breaking** with parameters μ and σ:

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_i} \quad \theta_h \overset{iid}{\sim} P_0$$

$$\pi_h = V_h \prod_{\ell<h} (1 - V_\ell)$$

$$V_h = \Phi(x_h), \quad \text{where } x_h \overset{iid}{\sim} N(\mu, \sigma^2)$$

Finite mixture models

- **Mixture of normals**

$$p(y) = \sum_{j=1}^{k} \pi_j \cdot \phi(y; \mu_j, \sigma_j^2)$$

- Use **allocation variables**: $I_i = j$ if y_i comes from $\phi(y; \mu_j, \sigma_j^2)$.
- Let $I = (I_1, ..., I_n)'$ and $n_j = \sum_{i=1}^{n} (I_i = j)$.
- **Gibbs sampling** algorithm:
  - $\pi_1, ..., \pi_k \mid I, y \sim Dirichlet(a_1 + n_1, a_2 + n_2, ..., a_k + n_k)$
  - $\sigma_j^2 \mid I, y \sim Inv\text{-}\chi^2$ and $\mu_j | I, \sigma_j^2, y \sim N$ for $j = 1, ..., k$.
  - $I_i \mid \pi, \mu, \sigma^2, y \sim Multinomial(\omega_{i,1}, ..., \omega_{i,k})$, $i = 1, ..., n$, where

$$\omega_{i,j} = \frac{\pi_j \cdot \phi(y_i; \mu_j, \sigma_j^2)}{\sum_{q=1}^{k} \pi_q \cdot \phi(y_i; \mu_q, \sigma_q^2)}.$$

Infinite mixture models - DP mixtures

- **General mixture formulation**

$$f(y|P) = \int \mathcal{K}(y|\theta) dP(\theta)$$

  where $\mathcal{K}(y|\theta)$ is a kernel and $P(\theta)$ is a **mixing measure**.

- Example 1: **Student**-t, $t_\nu(\mu, \sigma^2)$. $\mathcal{K}(y|\theta) = \phi(y|\mu, \lambda)$ where μ is fixed, $\theta = \lambda$ and $P(\theta)$ is the $Inv - \chi^2(\nu, \sigma^2)$ distribution.
- Example 2: **Finite mixture of normals**. $\phi(y|\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. $P(\theta)$ is a discrete distr. with $\Pr\left[\theta = (\mu_j, \sigma_j^2)\right] = \pi_j$, for $j = 1, ..., k$.
- Example 3: $P \sim DP(\alpha P_0)$ yields the **infinite mixture**

$$f(y) = \sum_{h=1}^{\infty} \pi_h \mathcal{K}(y|\theta_h^*), \qquad \pi \sim \text{Stick}(\alpha).$$

DP mixture is like a finite mixture with large k

- In infinite mixtures **every observation has its own parameter** θ_i

$$y_i \sim \mathcal{K}(\theta_i)$$

- DP is almost surely discrete $\Rightarrow$ **ties**: some of the θ_i will have exactly the same values. **DP leads to clustering** of the θ_i.
- Each observation has **potentially** its own parameter θ_i, but that **parameter may be shared by other observations**.
- In finite mixture models each observation also has its "own" parameter

$$y_i | I_i \sim \mathcal{K}(\theta_{I_i})$$

$$I_i | \pi \sim Multinomial(\pi_1, ..., \pi_k)$$

$$\theta_i \sim P_0$$

$$\pi \sim Dirichlet(\alpha/k, ..., \alpha/k)$$

- Neal (2000) shows that this finite mixture model approaches the DP mixture when $k \to \infty$.

Marginalizing out P from a DP - Polya scheme

- **Hierarchical representation of DP mixtures**

$$y_i \sim \mathcal{K}(\theta_i), \qquad \theta_i \sim P \qquad P \sim DP(\alpha P_0)$$

- We can actually **marginalize out** P to obtain the **Polya scheme**

$$p(\theta_i|\theta_1,...,\theta_{i-1}) \sim \left(\frac{\alpha}{\alpha+i-1}\right) P_0(\theta_i) + \left(\frac{1}{\alpha+i-1}\right) \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- So $p(\theta_i|\theta_1,...,\theta_{i-1})$ is a mixture of the base measure P_0 and point masses at the previously "drawn" θ-values.

- Way to think about the scary 'Marginalizing out P': integrate out π in the finite mixture model and let $k \to \infty$. [Neal, 2000].

DPs and the Chinese restaurant process

- The so called **Polya scheme**:

$$p(\theta_i|\theta_1, ..., \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \left(\frac{1}{\alpha + i - 1}\right) \sum_{j=1}^{i-1} \delta_{\theta_j}$$

- **Chinese restaurant process**:
  - first customer sits at empty table and obtains the dish θ_1^* from P_0.
  - second customer
    - sits at first customer's table with probability $\frac{1}{1+\alpha}$ and has dish θ_1^* or
    - sits at a new table with probability $\frac{\alpha}{1+\alpha}$ and has dish $\theta_2^* \sim P_0$.
  - $\vdots$
  - the ith customer
    - sits at table with dish θ_j^* with a probability proportional to n_j, the number of customers sitting at table j or
    - sits at a new table with probability proportional to α.

Gibbs sampling DP mixtures - marginalizing P

- Similar to Gibbs sampling for finite mixtures. Data augmentation with mixture component indicators I_i.

1. **Update component allocation** for ith observation y_i by sampling from multinomial

$$\Pr(I_i = j|\cdot) \propto \begin{cases} n_j^{(-i)} \mathcal{K}(y_i|\theta_j^*) & \text{for } j = 1, ..., k^{(-i)} \\ \alpha \int \mathcal{K}(y_i|\theta) dP_0(\theta) & \text{for } j = k^{(-i)} + 1 \end{cases}.$$

2. **Update** the unique **parameter values** θ^* by sampling from

$$p(\theta_j^*|\cdot) \propto P_0(\theta_c^*) \prod_{i:I_j=j} \mathcal{K}(y_i|\theta_j^*)$$

- Note that, unlike finite mixtures, the I_i **are not independent** conditional on θ^*. This because we have marginalized out P. They have to be sampled **sequentially**.

Gibbs sampling for truncated DP mixtures

- Set upper bound N for the number of components. Approximate DP mixture with $\pi_h = 0$ for $h = N+1, \ldots$
- Posterior samping for infinite mixtures is now very similar to finite mixture. The I_i can be sampled independently.

1. Update component allocation for ith observation y_i by sampling from multinomial
$$\Pr(I_i = j|\cdot) \propto \pi_j \mathcal{K}(y_i|\theta_j^*) \quad \text{for } j = 1, 2, \ldots, N.$$
2. Update the stick-breaking weights [recall: $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$]
$$V_j|\cdot \sim Beta\left(1 + n_j, \alpha + \sum_{q=j+1}^{N} n_q\right) \qquad \text{for } j = 1, \ldots, N-1.$$
3. Update the unique parameter values $\theta_1^*, \ldots \theta_N^*$ by sampling just like in the finite mixture model. Sample θ^* from prior $P_0(\theta)$ for empty clusters.

MCMC FOR DP MIXTURES

- Let's look at the updating step:

$$\Pr(I_i = j|\cdot) \propto \begin{cases} n_j^{(-i)} \mathcal{K}(y_i|\theta_c^*) & \text{for } j = 1, ..., k^{(-i)} \\ \alpha \int \mathcal{K}(y_i|\theta) dP_0(\theta) & \text{for } j = k^{(-i)} + 1 \end{cases}.$$

- **A customer chooses table based on**:
  - the number of existing customers at the tables (with imaginary α customers at a new table)
  - how compatible the taste of the customer (y_i) is to the different dishes served at occupied tables (θ_c^*)
  - how compatible the taste of the customer (y_i) is to the different dishes that *may* be served at a new table.
  - A $P_0(\theta)$ with large variance is equivalent to an very experimental cook. You never know what you get ...
- Hyperparameter α clearly matters for the number of clusters (tables), but so does P_0.
- Hyperparameter α can be learned from data. Just add updating step.
- P_0 may contain hyperparameters (e.g. $P_0 = N(\mu, \sigma^2)$). Just add updating steps for those.

Mixture of multivariate regressions - Model

- The response vector $\mathbf{y}$ is p-dim. Covariates $\mathbf{x}$ is q-dim.
- The model is of the form

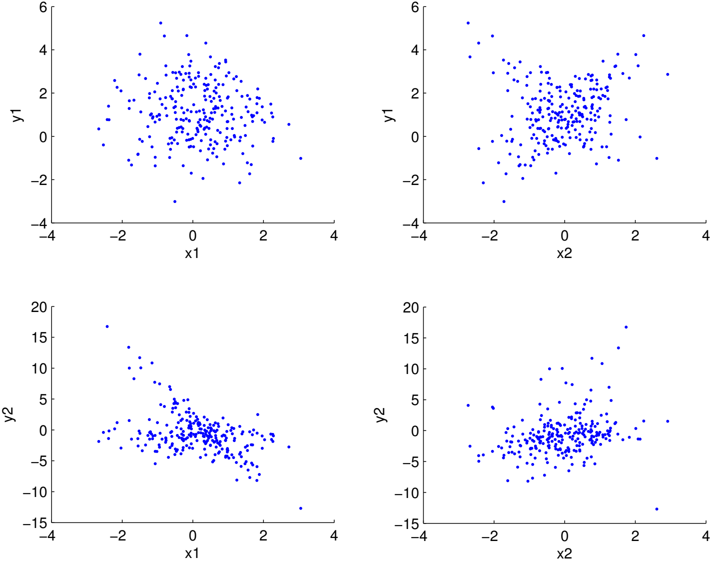
$$p(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{\infty} \pi_j \cdot N(\mathbf{y}_i|\mathbf{B}_j\mathbf{x}_i, \Sigma_j)$$

- Each component in the mixture is a Gaussian multivariate regression with its own regression coefficient and covariance matrix:

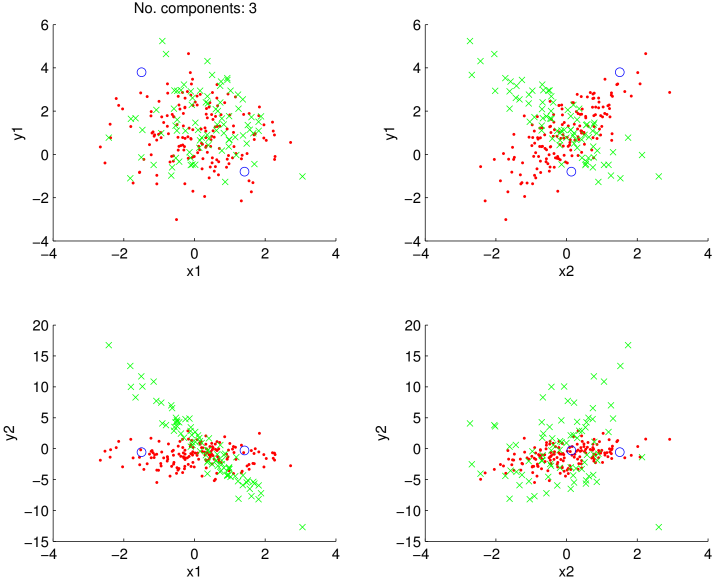
$$\underset{p\times 1}{\mathbf{y}_i} = \underset{p\times q}{\mathbf{B}_j}\underset{q\times 1}{\mathbf{x}_i} + \underset{p\times 1}{\varepsilon_i}, \ \varepsilon_i \overset{iid}{\sim} N\left(0, \Sigma_j\right)$$

- The mixture weights follow a DP stick prior $\pi \sim Stick(\alpha)$.

Mixture of multivariate regressions - Data



Mixture of multivariate regressions - DPM



Mixture of multivariate regressions - DPM