Dirichlet Process

Sara Wade
University of Cambridge

Charles University
8-19 April 2013, Prague

Categorical and multinomial distributions

Consider a discrete random variable X taking one of k possible outcomes. Among n independent and identical trials, let $n_j = \sum_{i=1}^{n} \mathbf{1}(x_i = j)$. Note that $n = \sum_{j=1}^{k} n_j$.

The distribution of X_i is given by the categorical distribution, parametrized by $p = (p_1, \ldots, p_k)$ such that $\sum p_j = 1$, where

$$p(x|p) = p_1^{\mathbf{1}(x=1)} * \ldots * p_k^{\mathbf{1}(x=k)}.$$

The probability of observing counts $(n_1, \ldots, n_k)$ is given by the multinomial distribution, where

$$p(n_1, \ldots, n_k|p) = \frac{n!}{n_1! \ldots n_k!} \prod_{j=1}^{k} p_j^{n_j}.$$

Ex. $(n_1, \ldots, n_k)$ is the frequency of words in a text.

Dirichlet distribution

The Dirichlet distribution is defined on $S_k = \{(p_1, \ldots, p_k) : p_j \geq 0, \sum_{j=1}^k p_j = 1\}$ with density

$$p((p_1, \ldots, p_k)|(\alpha_1, \ldots, \alpha_k)) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k p_j^{\alpha_j - 1}.$$

It is the **conjugate prior** to the multinomial likelihood.

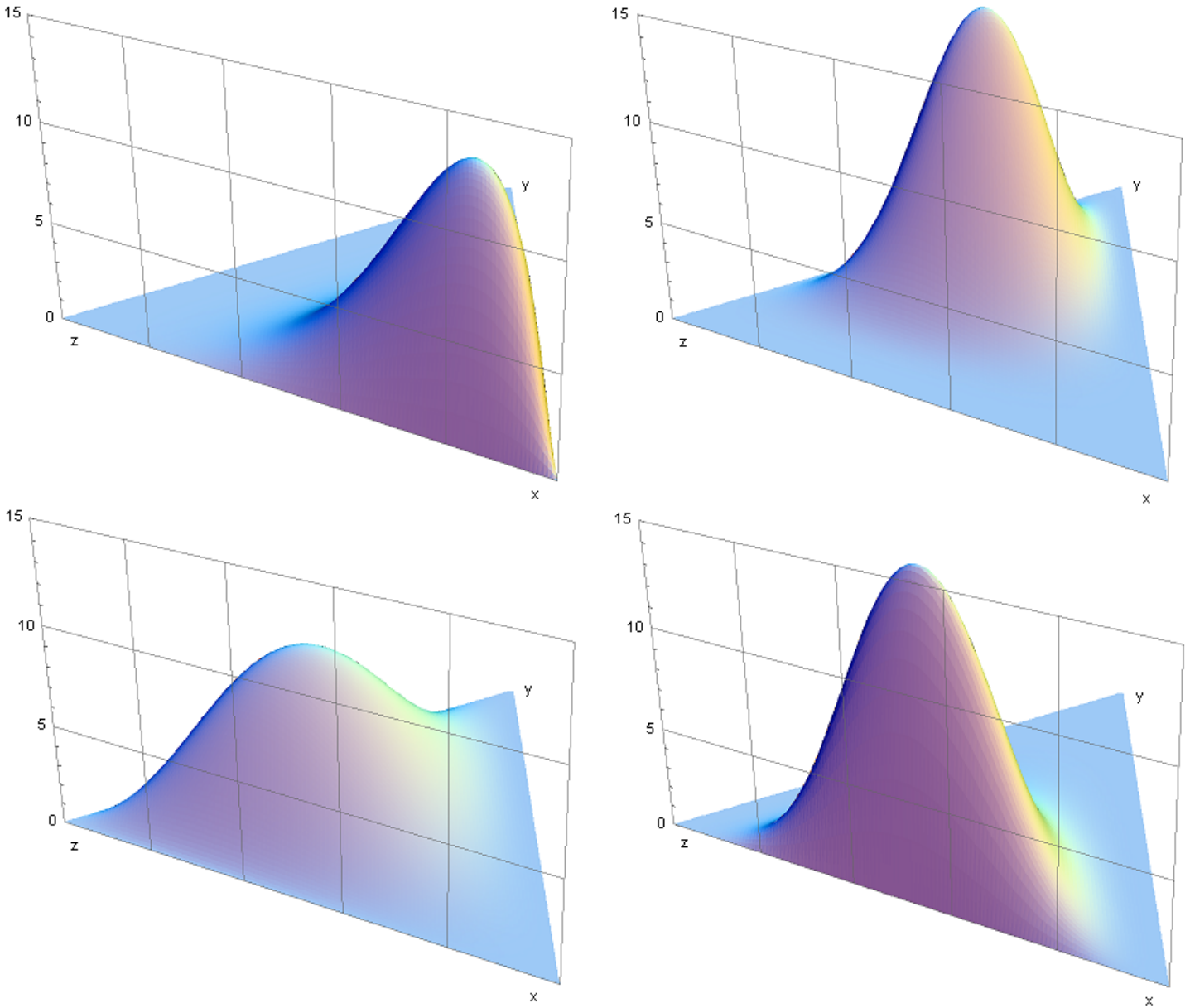
Parameters: $(\alpha_1, \ldots, \alpha_k)$ such that $\alpha_j \geq 0$, are often reparametrized as

$$\alpha = \sum_{j=1}^k \alpha_j; \quad p_0 = (p_{0\,1}, \ldots, p_{0\,k}) = \left(\frac{\alpha_1}{\alpha}, \ldots, \frac{\alpha_k}{\alpha}\right).$$

Properties:

- $\mathrm{E}[p_j] = p_{0\,j} \leftarrow$ **prior guess**.
- $\mathrm{V}(p_j) = \frac{p_{0\,j}(1-p_{0\,j})}{\alpha+1} \leftarrow \alpha$ **controls the variability**.

Dirichlet densities from Wikipedia



Connections with other distributions

1. if $z_j \overset{ind}{\sim} \text{Gam}(\alpha_i, 1)$, then

$$(p_1, \ldots, p_k) \overset{d}{=} \left(\frac{z_1}{\sum_{j=1}^k z_j}, \ldots, \frac{z_k}{\sum_{j=1}^k z_j} \right).$$

   This property is used to simulate from a Dirichlet distribution.

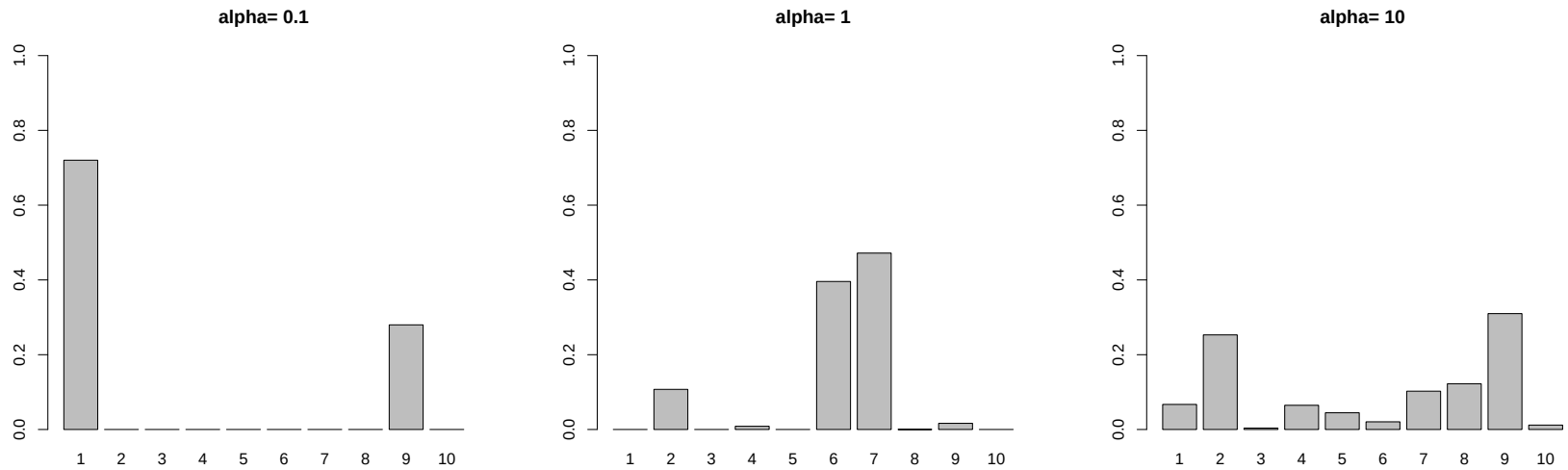
2. if $v_j \overset{ind}{\sim} \text{Beta}(\alpha_j, \sum_{j'>j} \alpha_{j'})$ for $j = 1, \ldots, k-1$ and v_k is degenerate at 1, then

$$(p_1, \ldots, p_k) \overset{d}{=} \left(v_1, v_2(1-v_1) \ldots, v_k \prod_{j<k} (1-v_j) \right).$$

   Note that $v_j = \frac{p_j}{1-\sum_{j'<j} p_{j'}}$.

Symmetric Dirichlet distribution

The symmetric Dirichlet distribution is defined with $p_{0\,j} = \frac{1}{k}$ for $j = 1, \ldots, k$.



Densities p drawn at random from a symmetric Dirichlet distribution with various precision parameters.

Posterior and Predictive

Categorical model:

$$X_i|p \overset{iid}{\sim} \text{Cat}(p).$$

Dirichlet prior:

$$p \sim \text{Dir}(\alpha p_0).$$

$\rightarrow$ Leads to a Dirichlet posterior

$$p|\mathbf{x} \sim \text{Dir}(\hat{\alpha}\hat{p}),$$

where

$$\hat{\alpha} = \alpha + n; \quad \hat{p_j} = \hat{\alpha}^{-1}(\alpha p_{0\,j} + n_j).$$

$\rightarrow$ Leads to a Categorical predictive with

$$p(X_{n+1} = j|\mathbf{x}) = \hat{p}_j.$$

Pólya urn scheme

The **Pólya urn scheme** describes the distribution of a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ taking values in $\{1,\ldots,k\}$.

Consider an urn with $\alpha p_{0\,j}$ balls of color j for $j = 1,\ldots,k$. A ball is drawn from the urn and replaced along with another ball of the same color. The random variable X_n is set to j if the n^{th} ball drawn is of color j.

Formally, the law of $\{X_n\}_{n\in\mathbb{N}}$ is given by

- $P(X_1 = j) = p_{0\,j}$,
- $P(X_{n+1} = j|\mathbf{x}) = \frac{\alpha p_{0\,j} + n_j}{\alpha + n}$ for $n > 1$.

Exchangeability and De Finetti's Theorem

The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ taking values in $\{1,\ldots,k\}$ is **exchangeable** if for any n and permutation π of $\{1,\ldots,n\}$

$$P(X_1 = j_1,\ldots,X_n = j_n) = P(X_{\pi(1)} = j_1,\ldots,X_{\pi(n)} = j_n),$$

for any $j_i \in \{1,\ldots,k\}$.

Theorem (De Finetti's Theorem)

A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ taking values in $\{1,\ldots,k\}$ is **exchangeable** *if and only if there exists a unique probability measure Q on S_k such that for any n and measurable sets any $j_i \in \{1,\ldots,k\}$,*

$$P(X_1 = j_1,\ldots,X_n = j_n) = \int_{S_k} \prod_{i=1}^{n} p_{j_i} dQ(p).$$

Pólya urn scheme and the Dirichlet distribution

- If the distribution of $\{X_n\}_{n\in\mathbb{N}}$ is described by the Pólya urn scheme, then $\{X_n\}_{n\in\mathbb{N}}$ is exchangeable.
- If $X_i|p$ have categorical distribution and $p \sim \text{Dir}(\alpha p_0)$, then the marginal distribution of $\{X_n\}_{n\in\mathbb{N}}$ is described by the Pólya urn scheme.

The distribution of $\{X_n\}_{n\in\mathbb{N}}$ is described by the Pólya urn scheme if and only if $X_i|p \overset{iid}{\sim} \text{Cat}(p)$ and $p \sim \text{Dir}(\alpha p_0)$.

Dirichlet Process

The Dirichlet process is an extension of the Dirichlet distribution on the space of probability measures on $\{1, \ldots, k\}$ to the space of probability measures on a complete and separable metric space $\mathcal{X}$.

Let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$, equipped with the Borel σ-algebra under weak convergence.

Definition

P has a Dirichlet process prior with parameters $\alpha > 0$ and $P_0 \in \mathcal{P}(\mathcal{X})$, denoted $\text{DP}(\alpha P_0)$, if for any finite measurable partition $(B_1, \ldots, B_m)$,

$$(P(B_1), \ldots, P(B_m)) \sim \text{Dir}(\alpha P_0(B_1), \ldots, \alpha P_0(B_m)).$$

Parameters:

- the base measure P_0 is the **prior guess**, $\text{E}[P(B)] = P_0(B)$,
- the precision parameter α **controls the variability**, $\text{V}(P(B)) = \frac{P_0(B)(1-P_0(B))}{\alpha+1}$.

Existence of the DP

Marginal property of the Dirichlet distribution: Let $B_1, \ldots, B_m$ be a partition $\{1, \ldots, k\}$ and $p(B_i) = \sum_{j \in B_i} p_j$,

$$(p(B_1), \ldots, p(B_m)) \sim \text{Dir}(\alpha p_0(B_1), \ldots, \alpha p_0(B_m)),$$

where $p_0(B_i) = \sum_{j \in B_i} p_{0\,j}$.

The marginal property of the Dirichlet distribution is a key property in showing existence of the Dirichlet process.

Stick-breaking construction

Theorem (Sethuraman (1994))

$P \sim DP(\alpha P_0)$ is characterized by the stick-breaking construction

$$P = \sum_{j=1}^{\infty} p_j \delta_{\theta_j},$$

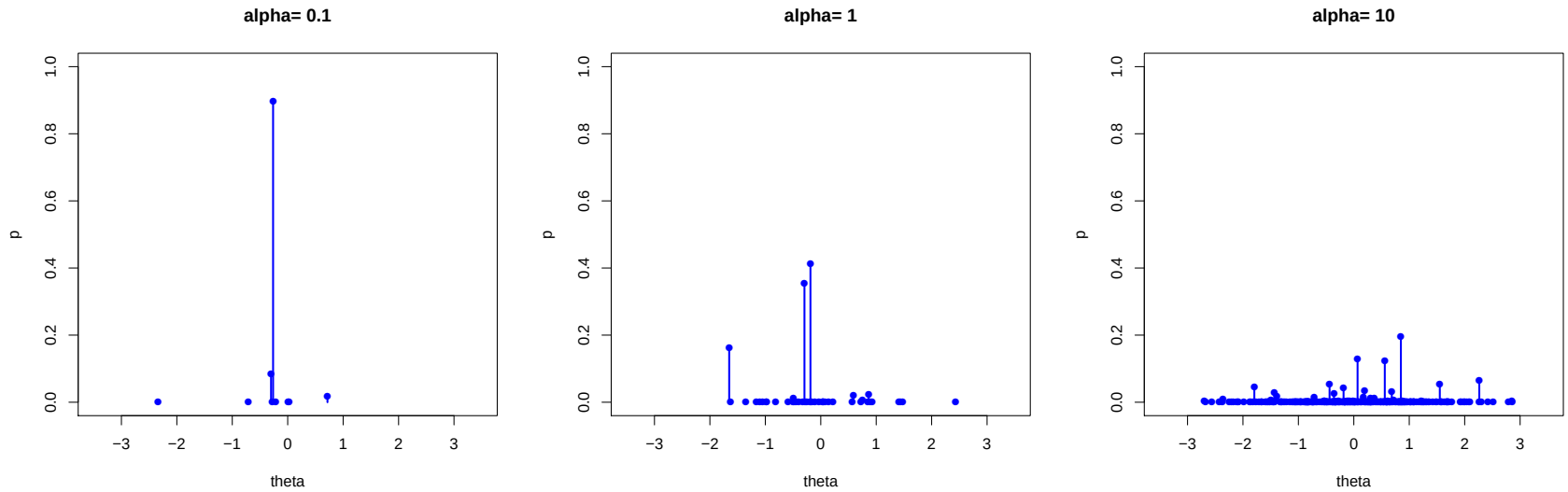
where $\theta_j \overset{iid}{\sim} P_0$,

$$p_1 = v_1; \quad p_j = v_j \prod_{j'<j} (1 - v_{j'}),$$

and $v_j \overset{iid}{\sim} Beta(1, \alpha)$ *independent of* (θ_j).

Notice: If $P \sim \text{DP}(\alpha P_0)$, P is **discrete** a.s.

DP prior samples



Random draws of $P \sim \text{DP}(\alpha \text{N}(0, 1))$ with various precision parameters. Simulation is based on the stick-breaking construction.

Posterior and Predictive

Model:

$$X_i|P \overset{iid}{\sim} P.$$

Dirichlet process prior:

$$P \sim \mathsf{DP}(\alpha P_0).$$

$\rightarrow$ Leads to a Dirichlet process posterior

$$P|\mathbf{x} \sim \mathsf{DP}(\hat{\alpha}\hat{P}),$$

where

$$\hat{\alpha} = \alpha + n; \quad \hat{P} = \hat{\alpha}^{-1}(\alpha P_0 + \sum_{i=1}^{n} \delta_{x_i}).$$

$\rightarrow$ Predictive distribution is

$$P(X_{n+1} \in B|\mathbf{x}) = \hat{P}(B),$$

for any Borel set $B \subset \mathcal{X}$.

Blackwell and MacQueen urn scheme

The **Blackwell and MacQueen urn scheme** describes the distribution of a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ taking values in $\mathcal{X}$.

Consider an urn with α black balls.

- Step 1: a black ball is drawn from the urn, and once drawn, its true color is revealed as θ_1^* from P_0; it is replaced along with a black ball.
- Step $n+1$: a ball is drawn from the urn. If the ball is black, once drawn, its true color is revealed as $\theta_{k_n+1}^*$, and it is replaced along with a black ball. Otherwise, it is of color θ_j^* for $j = 1, \ldots, k_n$, and it is replaced along with another ball of the same color.

Here k_n denotes the number of black balls drawn among the first n draws. We set $X_n = \theta_j^*$ if the n^{th} ball drawn is color θ_j^*.

Formally, the law of $\{X_n\}_{n\in\mathbb{N}}$ is given by

- $P(X_1 \in B) = P_0(B)$,
- $P(X_{n+1} \in B|\mathbf{x}) = \frac{\alpha P_0(B)+\sum_{i=1}^n \delta_{x_i}(B)}{\alpha+n}$ for $n > 1$.

for any Borel set $B \subseteq \mathcal{X}$.

Exchangeability and De Finetti's Theorem

The sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is **exchangeable** if for any n and permutation π of $\{1,\ldots,n\}$

$$P(X_1 \in B_1,\ldots,X_n \in B_n) = P(X_{\pi(1)} \in B_1,\ldots,X_{\pi(n)} \in B_n),$$

for measurable sets $B_i \subseteq \mathcal{X}$.

Theorem (De Finetti's Theorem)

A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is **exchangeable** *if and only if there exists a unique probability measure Q on $\mathcal{P}(\mathcal{X})$ such that for any n and measurable sets $B_i \subseteq \mathcal{X}$,*

$$P(X_1 \in B_1,\ldots,X_n \in B_n) = \int_{\mathcal{P}(\mathcal{X})} \prod_{i=1}^{n} P(B_i) dQ(P).$$

B+M urn scheme and the DP

Theorem (Blackwell and MacQueen (1973))

The distribution of $\{X_n\}_{n\in\mathbb{N}}$ is described by the Blackwell and MacQueen urn scheme if and only if $X_i|P \overset{iid}{\sim} P$ and $P \sim DP(\alpha P_0)$.

Clustering

Since P is discrete a.s., there is a positive probability of ties among the sample $(x_1, \ldots, x_n)$.

Let k_n denote the number of unique values; $(\theta_1^*, \ldots, \theta_{k_n}^*)$ denote the unique values; and n_j denote the cluster sizes.

Assuming P_0 is non-atomic, from the B+M urn scheme, we have

$$x_1 = \theta_1^*,$$

$$x_{n+1} \mid \mathbf{x} = \begin{cases} \theta_{k_n+1}^* & \text{with prob. } \frac{\alpha}{\alpha+n} \\ \theta_j^* & \text{with prob. } \frac{n_j}{\alpha+n} \text{ for } j = 1, \ldots, k \end{cases},$$

where $\theta_j^* \overset{iid}{\sim} P_0$.

The sample $(x_1, \ldots, x_n)$ can be represented in terms of the unique values $(\theta_1^*, \ldots, \theta_{k_n}^*)$ and the random partition $(s_1, \ldots, s_n)$ where $s_i = j$ if $x_i = \theta_j^*$. The predictive distribution of $(s_1, \ldots, s_n)$ is described by the **Chinese restaurant process.**

DP Mixture Models

Mixture models offer flexible density estimation:

$$p(x|P) = \int K(x|\theta)dP(\theta),$$

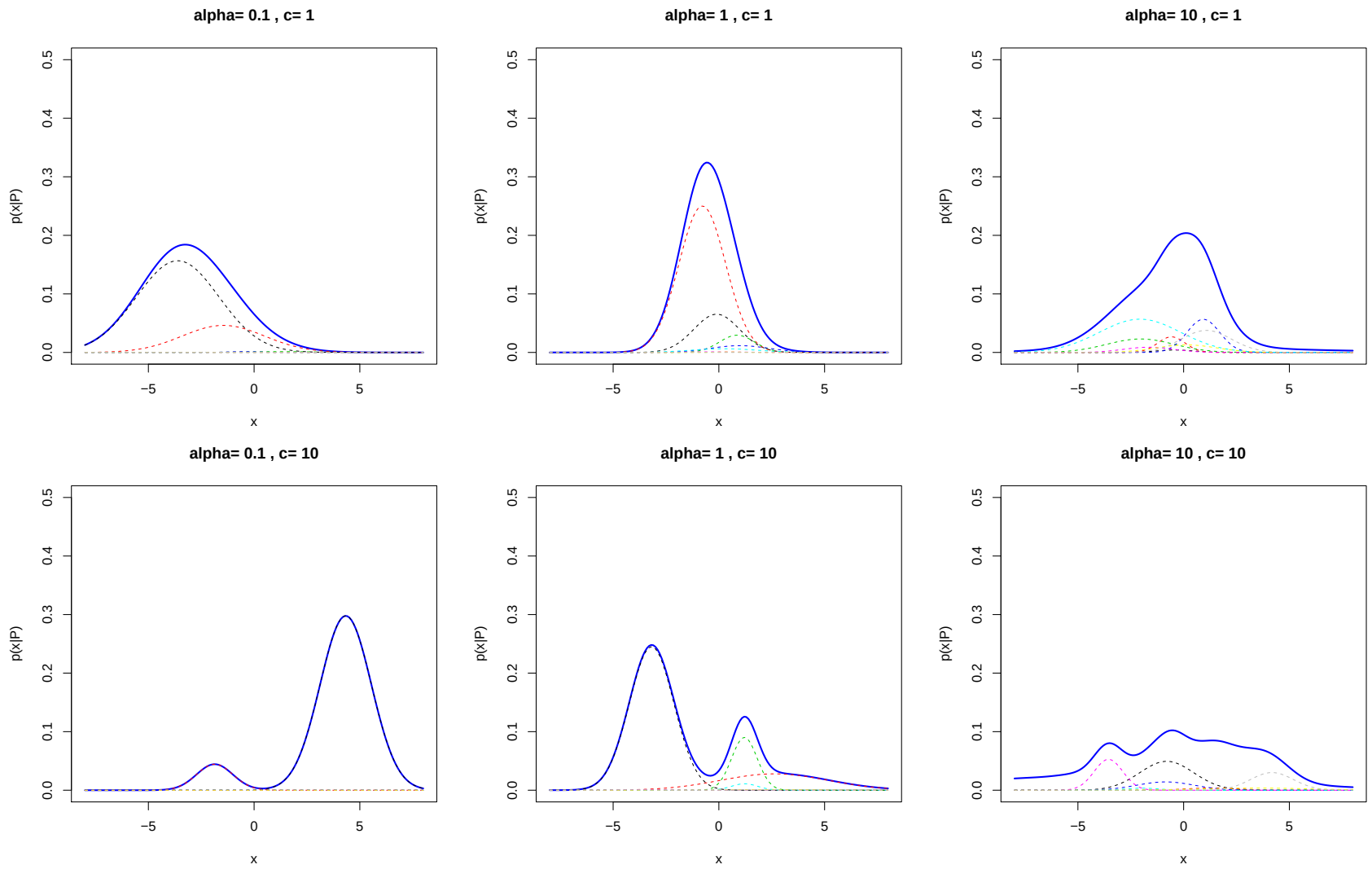
for some parametric density $K(x|\theta)$ (ex. $\mathsf{N}(x|\mu, \sigma^2)$).

In a Bayesian setting, we define a prior for P, ex.

$$P \sim \mathsf{DP}(\alpha P_0).$$

$$\Rightarrow p(x|P) = \sum_{j=1}^{\infty} p_j K(x|\theta_j).$$

DPM prior samples



Random draws of DP location-scale mixture of normals with base measure $\mathsf{N}(\mu|0, c\sigma^2)\mathsf{IG}(\sigma^2|1,1)$ with various values of α and c.

Inference in DPMs

The DPM model can be hierarchically defined as:

$$X_i|\theta_i \overset{ind}{\sim} K(x|\theta_i),$$
$$\theta_i|P \overset{iid}{\sim} P,$$
$$P \sim \text{DP}(\alpha P_0).$$

Marginal MCMC methods are based on the idea of marginalizing over P and carrying out posterior inference on $(\theta_1, \ldots, \theta_n)$ using Gibbs sampling based on the urn scheme characterization of the DP.

Other methods include **truncation**, **slice sampling**, and **retrospective sampling**.

Marginal Inference in DPMs

In, **marginal** MCMC methods the parameters $(\theta_1, \ldots, \theta_n)$ are represented as $s = (s_1, \ldots, s_n)$ and $\theta^* = (\theta_1^*, \ldots, \theta_{k_n}^*)$.

The algorithm then proceeds by

- for $i = 1, \ldots, n$, sample $s_i | \mathbf{x}, s^{-i}, \theta^*$ where

$$p(s_i = j | \mathbf{x}, s^{-i}, \theta^*) = \begin{cases} \frac{1}{Z}\alpha \int K(x_i|\theta) dP_0(\theta) & \text{for } j = k_n + 1 \\ \frac{1}{Z} n_j K(x_i|\theta_j^*) & \text{for } j = 1, \ldots, k_n \end{cases}.$$

- sample $\theta^* | \mathbf{x}, s$ where

$$p(\theta^* | \mathbf{x}, s) = \prod_{j=1}^{k_n} p(\theta_j^* | \mathbf{x}_j),$$

for $\mathbf{x}_j = (x_i)_{i:s_i=j}$, and

$$p(\theta_j^* | \mathbf{x}_j) \propto P_0(d\theta_j^*) \prod_{i:i=s_j} K(x_i|\theta_j^*).$$

Dependent Dirichlet Process

Dependent Dirichlet process priors define a distribution over a collection of random probability measures $\{P_x\}_{x \in \mathcal{X}}$ such that the P_x's are dependent and marginally P_x is a Dirichlet process.

If the input x is discrete and categorical, examples include

- Hierarchical DP (Teh et al. 2006): $P_m|P \overset{iid}{\sim} \text{DP}(\alpha P)$ for $m = 1, \ldots, M$; $P \sim \text{DP}(\beta P_0)$.
- Nested DP (Rodriguez and Dunson 2011): $P_m|Q \overset{iid}{\sim} Q$ for $m = 1, \ldots, M$; $Q \sim \text{DP}(\alpha \text{DP}(\beta P_0))$.

Dependent Dirichlet Process

MacEachern's (1999) general class of dependent Dirichlet processes are defined based on the stick-breaking representation:

$$P_x = \sum_{j=1}^{\infty} p_j(x)\delta_{\theta_j(x)},$$

where $\theta_j(x)$ are independent stochastic processes (ex. $\theta_j(x) \sim \text{GP}(0, k(x, x'))$), and

$$p_1(x) = v_1(x); \quad p_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)) \text{ for } j > 1,$$

for independent stochastic processes $v_j(x)$ such that marginally $v_j(x) \sim \text{Beta}(1, \alpha(x))$.

References

- Ghosh, J.K. and Ramamoorthi, R.V. (2003). Bayesian nonparametrics. *Springer Series in Statistics*.
- Rasmussen, C.E. and Ghahramani, Z. (2013). Machine learning course. `http://mlg.eng.cam.ac.uk/teaching/4f13/1213`