**10-708: Probabilistic Graphical Models 10-708, Spring 2014**

# 19 : Bayesian Nonparametrics: Dirichlet Processes

*Lecturer: Eric P. Xing* *Scribes: Carl Malings, Jingkun Gao*

## 1 Introduction

In parametric modeling, it is assumed that data can be represented by models using a fixed, finite number of parameters. Examples of parametric models include clusters of $K$ Gaussians and polynomial regression models. In many problems, determining the number of parameters a priori is difficult; for example, selecting the number of clusters in a cluster model, the number of segments in an image segmentation problem, the number of chains in a hidden Markov model, or the number of topics in a topic modelling problem before the data is seen can be problematic.

In nonparametric modeling, the number of parameters is not fixed, and often grows with the sample size. Kernel density estimation is an example of a nonparametric model. In Bayesian nonparametrics, the number of parameters is itself considered to be a random variable. One example is to do clustering with $k$-means (or mixture of Gassuians) while the number of clusters $k$ is unknown. Bayesian inference addresses this problem by treating $k$ itself as a random variable. A prior is defined over an infinite dimensional model space, and inference is done to select the number of parameters. Such models have infinite capacity, in that they include an infinite number of parameters a priori; however, given finite data, only a finite set of these parameters will be used. Unused parameters will be integrated out.

### 1.1 Mixture Model Example

An example of a parametric model is the mixture model of $K$ Gaussians with finite number of parameters:

$$p(x_1, \cdots, x_N | \pi, \{\mu_k\}, \{\Sigma_k\}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

If we were to adopt a Bayesian approach, each parameter would be given a prior distribution and integrated out:

$$p(x_1, \cdots, x_N) = \int \cdots \int \left( \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right) p(\pi) p(\mu_{1:K}) p(\Sigma_{1:K}) d\pi d\mu_{1:K} d\Sigma_{1:K}$$

It is best to choose conjugate prior distributions to simplify posterior inference. For the Gaussian parameters, the Gaussian and inverse Wishart distributions are the conjugate distributions for the mean and co-variance respectively. For the mixture weights, the conjugate is the Dirichlet distribution.

## 2 The Dirichlet Distribution

The Dirichlet distribution is a distribution over the $(K-1)$-dimensional simplex; that is, it is a distribution over the relative values of $K$ components, whose sum is restricted to be 1. It is parameterized by a $K$-

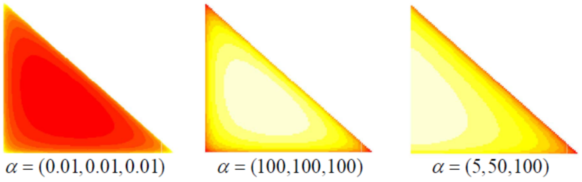

Figure 1: Density of the 3-component Dirichlet distribution for different parameter settings. Red indicates higher density.

dimensional vector $(\alpha_1, \cdots, \alpha_K)$, where $\alpha_k \geq 0 \; \forall \; k$ and $\sum_k \alpha_k > 0$. Its distribution is given by:

$$\pi = (\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

If $\pi \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ then $\pi_k \geq 0 \; \forall \; k$ and $\sum_{k=1}^K \pi_k = 1$. The expectation of the distribution is:

$$\mathbb{E}\left[(\pi_1, \cdots, \pi_K)\right] = \frac{(\alpha_1, \cdots, \alpha_k)}{\sum_k \alpha_k}$$

Figure 1 shows how the density of a Dirichlet distribution over 3 components varies for different settings of its scaling parameters $\alpha$. Note that as the parameter values become larger, the distribution becomes more concentrated at the extremes (i.e. it is more likely that one component take on value 1 and the rest value 0). Furthermore, different values for the parameters can skew the distribution.

## 2.1 Conjugacy with the Multinomial Distribution

It can be shown that the Dirichlet distribution is the conjugate of the Multinomial distribution. If $\pi \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ and $x_n \sim \text{Multinomial}(\pi)$ are *iid* samples, then:

$$\begin{aligned} p(\pi | x_1, \cdots, x_n) &\propto p(x_1, \cdots, x_n | \pi) p(\pi) \\ &= \left( \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \prod_{k=1}^K \pi_k^{\alpha_k - 1} \right) \left( \frac{n!}{m_1! \cdots m_K!} \pi_1^{m_1} \cdots \pi_K^{m_K} \right) \\ &\propto \frac{\prod_{k=1}^K \Gamma(\alpha_k + m_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + m_k\right)} \prod_{k=1}^K \pi_k^{\alpha_k + m_k - 1} \\ &= \text{Dirichlet}(\alpha_1 + m_1, \cdots, \alpha_K + m_K) \end{aligned}$$

Where $m_k$ represent the counts of instances of $x_n = k$ in the data set.

The Dirichlet distribution can be viewed as a distribution over finite-dimensional distributions; that is, it is a distribution over parameters for the Multinomial distribution, where each sample from the Dirichlet distribution can be regarded as a Multinomial distribution. Furthermore, we can associate each component with a set of parameters. In the case of a Gaussian mixture model, these parameters would be the mean and covariance of each cluster. We would therefore define a prior distribution over Gaussians. In the Bayesian setting, these parameters are themselves random variables. In a Bayesian finite mixture model, we combine a Gaussian prior over the centroid locations of clusters with a Dirichlet prior over the cluster weights.

## 2.2 Other Properties of the Dirichlet Distribution

The Dirichlet distribution satisfies the coalesce rule:

$$(\pi_1 + \pi_2, \pi_3, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \cdots, \alpha_K)$$

This property can be proved by examining the relationship between the Gamma and Dirichlet distributions. If $\eta_k \sim \text{Gamma}(\alpha_k, 1)$ represent $K$ independent Gamma-distributed variables, then their normalized sum follows a Dirichlet distribution:

$$\frac{(\eta_1, \cdots, \eta_K)}{\sum_k \eta_k} \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$$

Furthermore, if $\eta_1 \sim \text{Gamma}(\alpha_1, 1)$ and $\eta_2 \sim \text{Gamma}(\alpha_2, 1)$, then by the properties of the Gamma distribution:

$$\eta_1 + \eta_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, 1)$$

Therefore, if $(\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$, then:

$$(\pi_1 + \pi_2, \pi_3, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \cdots, \alpha_K)$$

proving the coalesce rule, which allows us to reduce the dimensionality of the Dirichlet distribution.

The Dirichlet distribution also satisfies the expansion or combination rule, which allows us to increase the dimensionality of a Dirichlet distribution. Note that the Dirichlet distribution over the 1-dimensional simplex is simply the Beta distribution. Let $(\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ and $\theta \sim \text{Beta}(\alpha_1 b, \alpha_1(1-b))$ for $0 < b < 1$. Then one dimension of the Dirichlet distribution can be split into two dimensions as follows:

$$(\pi_1\theta, \pi_1(1-\theta), \pi_2, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b, \alpha_1(1-b), \alpha_2, \cdots, \alpha_K)$$

More generally, if $\theta \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \cdots, \alpha_1 b_N)$ and $\sum_i b_i = 1$, then:

$$(\pi_1\theta_1, \cdots, \pi_1\theta_N, \pi_2, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b_1, \cdots, \alpha_1 b_N, \alpha_2, \cdots, \alpha_K)$$

Finally, the Dirichlet distribution also satisfies the renormalization rule. If $(\pi_1, \cdots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_K)$ then:

$$\frac{(\pi_2, \cdots, \pi_K)}{\sum_{k=1}^K \pi_k} \sim \text{Dirichlet}(\alpha_2, \cdots, \alpha_K)$$

## 2.3 Constructing an Infinite-Dimensional Prior

In problems such as clustering, the number of clusters is not known a priori. When defining a prior for the mixture weights, we need a distribution that allows an infinite number of clusters, so that we will always have more clusters than we will need in any given problem. An infinite mixture model has the form:

$$p(x_n|\pi, \{\mu_k\}, \{\Sigma_k\}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

We would therefore like to use a prior that has properties like that of the Dirichlet distribution (such as conjugacy with the Multinomial), but is infinite-dimensional.

To define such a distribution, we consider the following scheme. Begin with a two-component Dirichlet distribution, with scaling parameter $\alpha$ divided equally between both components, for the sake of symmetry:

$$\pi^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$$

Then, split off components according to the expansion rule:

$$\theta_1^{(2)}, \theta_2^{(2)} \sim \text{Beta}(\frac{\alpha}{2} \times \frac{1}{2}, \frac{\alpha}{2} \times \frac{1}{2})$$

$$\pi^{(4)} = (\theta_1^{(2)}\pi_1^{(2)}, (1-\theta_1^{(2)})\pi_1^{(2)}, \theta_2^{(2)}\pi_2^{(2)}, (1-\theta_2^{(2)})\pi_2^{(2)}) \sim \text{Dirichlet}\left(\frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}\right)$$

We repeat this process, such that:

$$\pi^{(K)} \sim \text{Dirichlet}(\frac{\alpha}{K}, \cdots, \frac{\alpha}{K})$$

In the limit as $K$ goes to infinity, we will get a prior over an infinite-dimensional space. In practice, we will never need to use all of these components, only the finitely many components which are reflected in the data.

# 3 Dirichlet Process

The previous arguments motivate us to define the Dirichlet Process. Let the base measure $H$ be a distribution over some space $\Omega$ (for example, a Gaussian distribution over the real line). Let:

$$\pi \sim \lim_{K \to \infty} \text{Dirichlet}\left(\frac{\alpha}{K}, \cdots, \frac{\alpha}{K}\right)$$

For each point in this Dirichlet distribution, we associate a a draw from the base measure:

$$\theta_k \sim H \text{ for } k = 1, ..., \infty$$

Then:

$$G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

is an infinite discrete distribution over the continuous space $\Omega$. We write this as a Dirichlet Process:

$$G \sim \text{DP}(\alpha, H)$$

Samples from the Dirichlet Process are discrete. The point masses in the resulting distribution are called atoms; their positions in $\Omega$ are drawn from the base measure $H$, while their weights are drawn from an infinite-dimensional Dirichlet distribution. The concentration parameter $\alpha$ determines the distribution over atom weights; smaller values lead to sparser distributions, with larger weights on each atom.

A Dirichlet Process is a unique distribution over probability distributions on some space $\Omega$ such that for any finite partition $A_1, \cdots, A_K$ of $\Omega$, the total mass assigned to each partition is distributed according to:

$$(P(A_1), \cdots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_K))$$

Note that $H$ may be un-normalized. Furthermore, a CDF, $G$, on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition $(\phi_1, \phi_2, \cdots, \phi_m)$:

$$(G(\phi_1), \cdots, G(\phi_m)) \sim \text{Dirichlet}(\alpha G_0(\phi_1), \cdots, \alpha G_0(\phi_m))$$

where $G_0$ is the base measure and $\alpha$ is the scale parameter.

## 3.1 Conjugacy

Let $A_1, \cdots, A_k$ be a partition of $\Omega$, and let $H$ be a measure on $\Omega$. Let $P(A_k)$ be the mass assigned by $G \sim \mathrm{DP}(\alpha, H)$ to partition $A_k$. Then:

$$(P(A_1), \cdots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_K))$$

If we see an observation in the $J^{\text{th}}$ segment (or fraction), then:

$$(P(A_1), \cdots, P(A_j), \cdots, P(A_K)|X_1 \in A_j) \sim \text{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_j) + 1, \cdots, \alpha H(A_K))$$

Since this must be true for all possible partitions of $\Omega$, this is only possible if the posterior for $G$ is given by:

$$G|X_1 = x \sim \mathrm{DP}(\alpha + 1, \frac{\alpha H + \delta_x}{\alpha + 1})$$

## 3.2 Predictive Distribution

The Dirichlet distribution can be a prior for mixture models, thus the Dirichlet Process could be further used to cluster observations. A new data point can either join an existing cluster or start a new cluster. Assume $H$ is a continuous distribution on $\Omega$ and we have $\theta$ in $\Omega$ being parameters for modeling the observed data points. Once we have a first data point, we start a new cluster with a sampled parameter $\theta_1$. Now, we can split our parameter space in two: the singleton $\theta_1$, and everything else. Let $\pi_1$ be the atom at $\theta_1$. The combined mass of all the other atoms is $\pi_* = 1 - \pi_1$, and we have:

$$\text{prior} : (\pi_1, \pi_*) \sim \text{Dirichlet}(0, \alpha)$$

$$\text{posterior} : (\pi_1, \pi_*)|X_1 = \theta_1 \sim \text{Dirichlet}(1, \alpha)$$

If we integrate out $\pi_1$, we get:

$$\begin{aligned} P(X_2 = \theta_k|X_1 = \theta_1) &= \int P(X_2 = \theta_k|(\pi_1, \pi_*))P((\pi_1, \pi_*)|X_1 = \theta_1)\, d\pi_1 \\ &= \int \pi_k \text{Dirichlet}(1, \alpha)\, d\pi_1 \\ &= \mathbb{E}_{\text{Dirichlet}(1-\alpha)}[\pi_k] \\ &= \begin{cases} \frac{1}{1+\alpha} & \text{if } k = 1 \\ \frac{\alpha}{1+\alpha} & \text{for new k.} \end{cases} \end{aligned}$$

This basically tells that with probability $\frac{1}{1+\alpha}$, the parameter $\theta$ stays in the old cluster and with probability $\frac{\alpha}{1+\alpha}$ it starts a new cluster. Let's say we choose to start a new cluster and sample a new parameter $\theta_2 \sim H$. Let $\pi_2$ be the size of the atom at $\theta_2$, and we have:

$$\text{posterior} : (\pi_1, \pi_2, \pi_*)|X_1 = \theta_1, X_2 = \theta_2 \sim \text{Dirichlet}(1, 1, \alpha).$$

If we integrate out $\pi = (\pi_1, \pi_2, \pi_*)$, we get:

$$\begin{aligned} P(X_3 = \theta_k|X_1 = \theta_1, X_2 = \theta_2) &= \int P(X_3 = \theta_k|\pi)P(\pi|X_1 = \theta_1, X_2 = \theta_2)\, d\pi \\ &= \mathbb{E}_{\text{Dirichlet}(1,1,\alpha)}[\pi_k] \\ &= \begin{cases} \frac{1}{2+\alpha} & \text{if } k = 1 \\ \frac{1}{2+\alpha} & \text{if } k = 2 \\ \frac{\alpha}{2+\alpha} & \text{for new k.} \end{cases} \end{aligned}$$

In general, if $m_k$ is the number of times we have seen $X_i = k$, and $K$ is the total number of observed values:

$$\begin{aligned} P(X_{n+1} = \theta_k | X_1, \cdots, X_n) &= \int P(X_{n+1} = \theta_k | \pi) P(\pi | X_1, \cdots, X_n) \, d\pi \\ &= \mathbb{E}_{\text{Dirichlet}(m_1, \cdots, m_K, \alpha)}[\pi_k] \\ &= \begin{cases} \frac{m_k}{n+\alpha} & \text{if } k \leq K \\ \frac{\alpha}{n+\alpha} & \text{for new cluster.} \end{cases} \end{aligned}$$

which gives a simple closed-form predictive distribution for the next observation. Such a predictive distribution is especially useful for sampling and inference in Dirichlet Processes. Note that this distribution has a "rich-get-richer" property; clusters with more observations are more likely to have new observations. However, there is always the possibility of seeing a novel observation, with $\alpha$ controlling the tendency to initiate a new cluster.

### 3.3 Useful Metaphors

Several useful metaphors exist for helping to understand the Dirichlet Process.

**Pólya Urn Process:** Consider an urn with a black ball of mass $\alpha$. Iteratively sample balls from the urn with probability proportional to their mass. If the ball is black, return it to the urn, choose a previously unseen color, and add a unit mass ball of that color to the urn. If the ball is colored, return it an another unit mass ball of the same color to the urn.

**Chinese Restaurant Process:** Consider a Chinese restaurant with infinitely many tables. As customers enter, they may sit at an occupied table with probability proportionate to how many customers are already seated there, or they may sit at an unoccupied table with probability proportionate to $\alpha$. Also, at each table, a dish is selected and shared by the customers seated there; this is analogous to a draw $\theta$ from the base measure $H$. We can also see from this example that the distribution does not depend on the ordering in which the customers arrived; this is the property of exchangeability. This way, each customer can be treated indepdendently, as if they were the last to arrive; this is a useful property for Gibbs sampling in Dirichlet Processes.

**Stick-breaking Process:** Consider a stick of unit length. Iteratively sample a random variable $b_k$ from $\text{Beta}(1, \alpha)$ and break off a fraction of $b_k$ of the stick. This is the weight of the $k^{\text{th}}$ atom. Sample a location for this atom from the base measure $H$ and repeat. This will give an infinite number of atoms, but as the process continues, the weights of the atoms will decrease to be negligible. This motivates truncation approximations, in which the full Dirichlet Process is approximated using a finite number of clusters.

## 4 Graphical Model Representation of DP

Figure 2 presents two alternative graphical models of the Dirichlet Process.

In the Pólya Urn (or Chinese restaurant) construction, $\{x_i\}$ is the data point sampled from the centroid $\{\theta_i\}$, which is itself sampled from the infinite dimensional distribution which follows Dirichlet Process $G$. $G$ is defined by a base measure $G_0$ and scaling parameter $\alpha$.
In the stick-breaking construction, the data $\{x_i\}$ has an indicator function of the cluster assignments $\{y_i\}$ which follows a multinomial distribution $\pi$ decided by a stick-breaking process. The centroid of the cluster $\theta$ follows a base measure $G_0$, which could be a Gaussian.

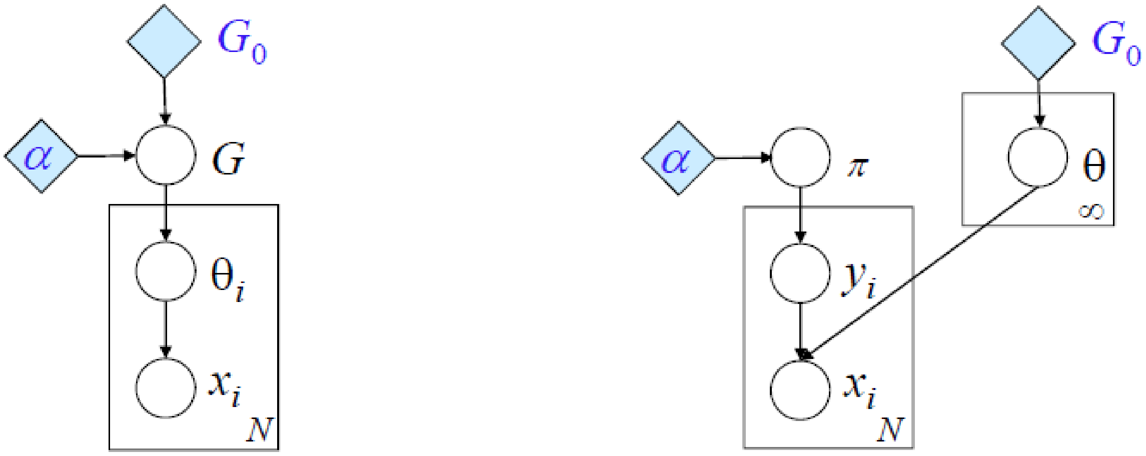

Figure 2: Two graphical models of the Dirichlet Process. The left model is derived from the Pólya Urn construction; the right model uses the stick-breaking construction.

# 5 Inference

The inference procedure tries to figure out what are the hidden associations of the points to the infinite number of parameters. We consider a Dirichlet Process mixture model, which is one step beyond the Dirichlet Process:

$$\begin{aligned} G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} &\sim \mathrm{DP}(\alpha, H) \\ \phi_n &\sim G \\ x_n &\sim f(\phi_n) \end{aligned}$$

where parameters $\phi_n$ from the Dirichlet Process distribution will be used as centroids to generate the actual samples $x_n$ in a mixture model. For example, $f$ could be a Gaussian centered at mean $\phi_n$.

## 5.1 Collapsed Sampler

The collapsed sampler integrates out $G$ to obtain a Chinese restaurant process. Since samples are exchangeable, we can treat any sampling point as the last point. Then:

$$p(z_n = k | x_n, z_{-n}, \phi_{1:K}) \propto \begin{cases} \frac{m_k}{n+\alpha} f(x_n|\phi_k) & k \leq K \\ \frac{\alpha}{n+\alpha} \int_\Omega f(x_n|\phi) H(d\phi) & k = K+1 \end{cases}$$

where $z_n$ is the cluster allocation of the $n^{\text{th}}$ data point, and $K$ is the total number of instantiated clusters. A potential problem with this sampling approach is that, since we are only sampling one data point at a time, mixing can be slow. Furthermore, if the likelihood is not conjugate, integrating out parameter values for new features can be difficult.

### 5.2 Blocked Gibbs Sampler

In the blocked Gibbs sampler, rather than integrating out $G$, we instantiate it. Although $G$ is infinite-dimensional, we can approximate it with a truncated stick-breaking process.

$$\begin{aligned} G^K &:= \sum_{k=1}^{K} \pi_k \delta_{\theta_k} \\ \pi_k &= b_k \prod_{j=1}^{k-1} (1 - b_j) \\ b_k &\sim \text{Beta}(1, \alpha) \text{ for } k = 1, ..., K-1 \\ b_K &= 1 \end{aligned}$$

We then sample cluster indicators and stick breaking variables as follows:

$$\begin{aligned} p(z_n = k|\text{rest}) &\propto \pi_k f(w_n|\theta_k) \\ b_k|\text{rest} &\sim \text{Beta}\left(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j\right) \end{aligned}$$

However, such a fixed truncation tends to introduce error.

### 5.3 Slice Sampler

The slice sampler induces a random truncation to the infinite $G$. Marginalizing over this random truncation, we would recover the full model. We do this by inducing a random variable $u_n$ for each data point:

$$u_n|\text{rest} \sim \text{Uniform}[0, \pi_{z_n}]$$

Conditioned on $u_n$ and $z_n$, $\pi_k$ can be sampled according to the block Gibbs sampler. We sample indicators according to:

$$p(z_n = k|\text{rest}) = I(\pi_k > u_n) f(x_n|\theta_k)$$

This scheme only represents a finite number of components $K$ such that:

$$1 - \sum_{k=1}^{K} \pi_k < min(u_n)$$

## 6 Summary

Bayesian nonparametrics in general involve ways of defining infinite-dimensional priors over parameters. An example is the Dirichlet Process described above. This process is motivated by moving from finite to infinite mixture models. More complex constructions, such as hierarchical processes and spatial or temporal sequences are also possible. Explicit expressions for the structures of distributions defined by these processes can motivate approximate schemes.