Dirichlet Processes: Tutorial and Practical Course

Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

August 2007 / MLSS

Dirichlet Processes

- Dirichlet processes (DPs) are a class of Bayesian nonparametric models.
- Dirichlet processes are used for:
  - Density estimation.
  - Semiparametric modelling.
  - Sidestepping model selection/averaging.
- I will give a tutorial on DPs, followed by a practical course on implementing DP mixture models in MATLAB.
- Prerequisites: understanding of the Bayesian paradigm (graphical models, mixture models, exponential families, Gaussian processes)—you should know these from Zoubin and Carl.
- Other tutorials on DPs:
  - Zoubin Gharamani, UAI 2005.
  - Michael Jordan, NIPS 2005.
  - Volker Tresp, ICML nonparametric Bayes workshop 2006.

Outline

Outline

Function Estimation

- Parametric function estimation (e.g. regression, classification)
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}, \mathbf{y} = \{y_1, y_2, \ldots\}$
  Model: $y_i = f(x_i|w) + \mathcal{N}(0, \sigma^2)$
- Prior over parameters

$$p(w)$$

- Posterior over parameters

$$p(w|\mathbf{x}, \mathbf{y}) = \frac{p(w)p(\mathbf{y}|\mathbf{x}, w)}{p(\mathbf{y}|\mathbf{x})}$$

- Prediction with posteriors

$$p(y_\star|x_\star, \mathbf{x}, \mathbf{y}) = \int p(y_\star|x_\star, w)p(w|\mathbf{x}, \mathbf{y})\, dw$$

Function Estimation

- Bayesian nonparametric function estimation with Gaussian processes
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}, \mathbf{y} = \{y_1, y_2, \ldots\}$
  Model: $y_i = f(x_i) + \mathcal{N}(0, \sigma^2)$
- Prior over functions

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

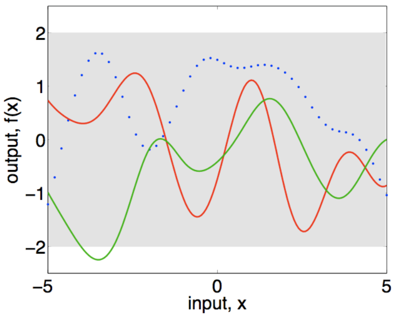
- Posterior over functions

$$p(f|\mathbf{x}, \mathbf{y}) = \frac{p(f)p(\mathbf{y}|\mathbf{x}, f)}{p(\mathbf{y}|\mathbf{x})}$$

- Prediction with posteriors

$$p(y_\star|x_\star, \mathbf{x}, \mathbf{y}) = \int p(y_\star|x_\star, f)p(f|\mathbf{x}, \mathbf{y})\, df$$

Function Estimation



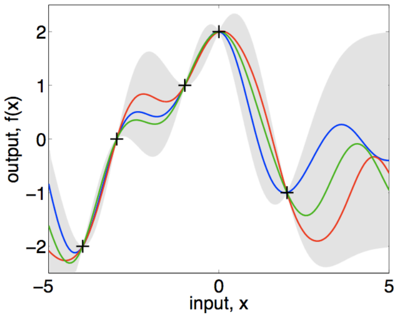


Figure from Carl's lecture.

Density Estimation

- Parametric density estimation (e.g. mixture models)

  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$

  Model: $x_i | w \sim F(\cdot|w)$

- Prior over parameters

$$p(w)$$

- Posterior over parameters

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|w)p(w|\mathbf{x})\,dw$$

Density Estimation

- Bayesian nonparametric density estimation with Dirichlet processes
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Model: $x_i \sim F$
- Prior over distributions

$$F \sim \mathrm{DP}(\alpha, H)$$

- Posterior over distributions

$$p(F|\mathbf{x}) = \frac{p(F)p(\mathbf{x}|F)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|F)p(F|\mathbf{x})\,dF = \int F'(x_\star)p(F|\mathbf{x})\,dF$$

- *Not quite correct; see later.*

Density Estimation

- Bayesian nonparametric density estimation with Dirichlet processes
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Model: $x_i \sim F$
- Prior over distributions

$$F \sim \mathrm{DP}(\alpha, H)$$

- Posterior over distributions

$$p(F|\mathbf{x}) = \frac{p(F)p(\mathbf{x}|F)}{p(\mathbf{x})}$$

- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|F)p(F|\mathbf{x})\,dF = \int F'(x_\star)p(F|\mathbf{x})\,dF$$

- *Not quite correct; see later.*

Density Estimation

- Bayesian nonparametric density estimation with Dirichlet processes
  Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Model: $x_i \sim F$
- Prior over distributions

$$F \sim \mathrm{DP}(\alpha, H)$$

- Posterior over distributions

$$p(F|\mathbf{x}) = \frac{p(F)p(\mathbf{x}|F)}{p(\mathbf{x})}$$

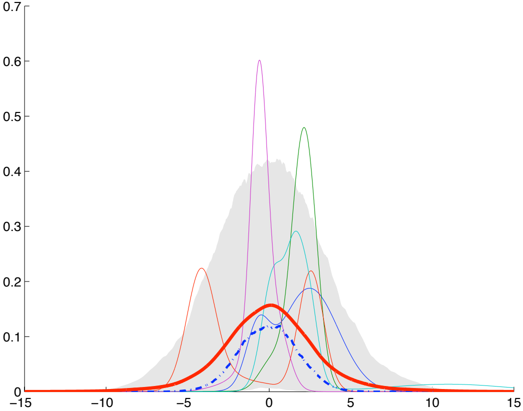
- Prediction with posteriors

$$p(x_\star|\mathbf{x}) = \int p(x_\star|F)p(F|\mathbf{x})\,dF = \int F'(x_\star)p(F|\mathbf{x})\,dF$$

- *Not quite correct; see later.*

Density Estimation

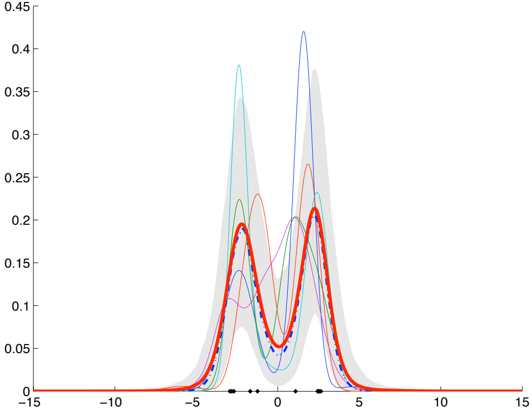
Prior:



Red: mean density. Blue: median density. Grey: 5-95 quantile.
Others: draws.

Density Estimation

Posterior:



Red: mean density. Blue: median density. Grey: 5-95 quantile.
Black: data. Others: draws.

Semiparametric Modelling

- Linear regression model for inferring effectiveness of new medical treatments.

$$y_{ij} = \beta^\top x_{ij} + b_i^\top z_{ij} + \epsilon_{ij}$$

  y_{ij} is outcome of jth trial on ith subject.

  x_{ij}, z_{ij} are predictors (treatment, dosage, age, health...).

  β are fixed-effects coefficients.

  b_i are random-effects subject-specific coefficients.

  ϵ_{ij} are noise terms.

- Care about inferring β. If x_{ij} is treatment, we want to determine $p(\beta > 0|\mathbf{x}, \mathbf{y})$.

Semiparametric Modelling

$$y_{ij} = \beta^\top x_{ij} + b_i^\top z_{ij} + \epsilon_{ij}$$

- Usually we assume Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Is this a sensible prior? Over-dispersion, skewness,...
- May be better to model noise nonparametrically,

$$\begin{aligned} \epsilon_{ij} &\sim F \\ F &\sim \mathrm{DP} \end{aligned}$$

- Also possible to model subject-specific random effects nonparametrically,

$$\begin{aligned} b_i &\sim G \\ G &\sim \mathrm{DP} \end{aligned}$$

Model Selection/Averaging

- Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Models: $p(\theta_k|M_k)$, $p(\mathbf{x}|\theta_k, M_k)$
- Marginal likelihood

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k)\,d\theta_k$$

- Model selection

$$M = \operatorname*{argmax}_{M_k} p(\mathbf{x}|M_k)$$

- Model averaging

$$p(x_\star|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)p(M_k|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)\frac{p(\mathbf{x}|M_k)p(M_k)}{p(\mathbf{x})}$$

- *But: is this computationally feasible?*

Model Selection/Averaging

- Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Models: $p(\theta_k|M_k)$, $p(\mathbf{x}|\theta_k, M_k)$
- Marginal likelihood

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k)\, d\theta_k$$

- Model selection

$$M = \operatorname*{argmax}_{M_k} p(\mathbf{x}|M_k)$$

- Model averaging

$$p(x_\star|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)p(M_k|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)\frac{p(\mathbf{x}|M_k)p(M_k)}{p(\mathbf{x})}$$

- *But: is this computationally feasible?*

Model Selection/Averaging

- Data: $\mathbf{x} = \{x_1, x_2, \ldots\}$
  Models: $p(\theta_k|M_k)$, $p(\mathbf{x}|\theta_k, M_k)$
- Marginal likelihood

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k)\,d\theta_k$$

- Model selection

$$M = \operatorname*{argmax}_{M_k} p(\mathbf{x}|M_k)$$

- Model averaging

$$p(x_\star|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)p(M_k|\mathbf{x}) = \sum_{M_k} p(x_\star|M_k)\frac{p(\mathbf{x}|M_k)p(M_k)}{p(\mathbf{x})}$$

- *But: is this computationally feasible?*

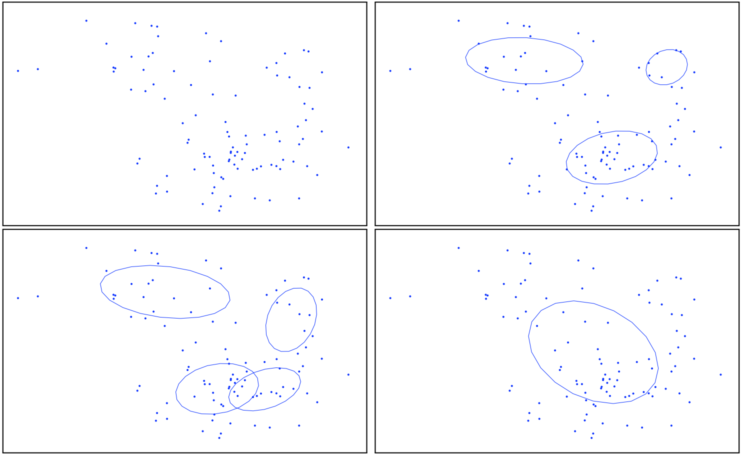
Model Selection/Averaging

- Marginal likelihood is usually extremely hard to compute.

$$p(\mathbf{x}|M_k) = \int p(\mathbf{x}|\theta_k, M_k)p(\theta_k|M_k)\, d\theta_k$$

- Model selection/averaging is to prevent underfitting and overfitting.
- But reasonable and proper Bayesian methods should not overfit [Rasmussen and Ghahramani 2001].
- Use a really large model M_∞ instead, and let the data speak for themselves.

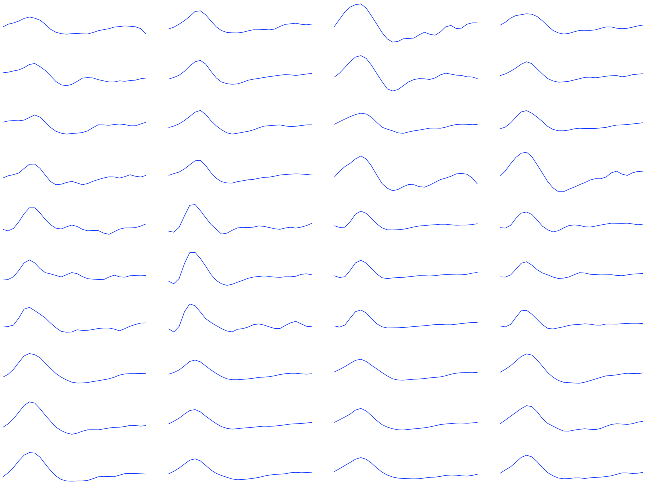
How many clusters are there?



Model Selection/Averaging

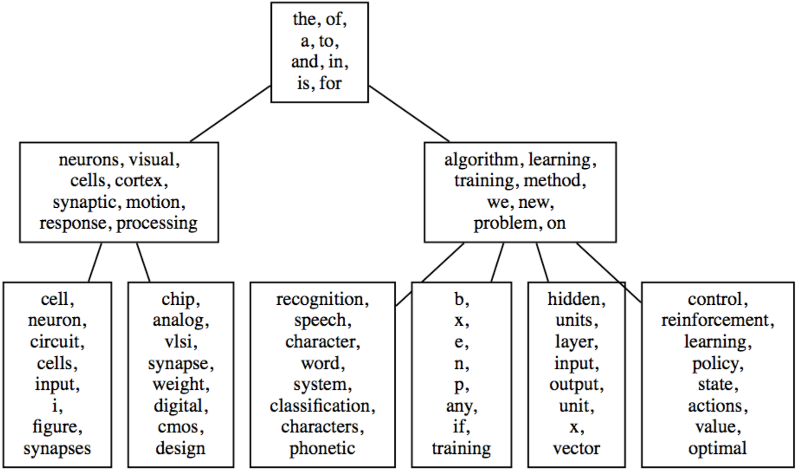
Spike Sorting

How many neurons are there?



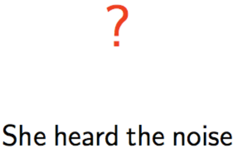
[Görür 2007, Wood et al. 2006]

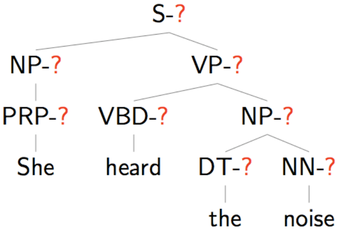
How many topics are there?



[Blei et al. 2004, Teh et al. 2006]

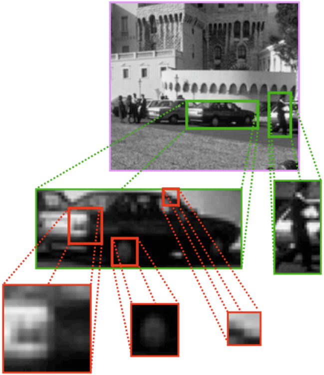
How many grammar symbols are there?





[Liang et al. 2007, Finkel et al. 2007]

How many objects, parts, features?



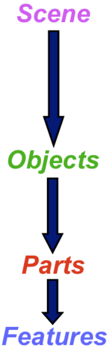


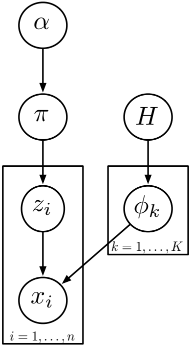
Figure from Sudderth. [Sudderth et al. 2007]

Outline

Finite Mixture Models

- A finite mixture model is defined as follows:

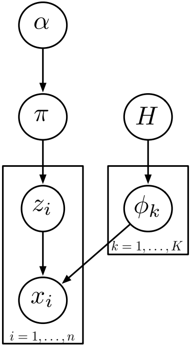
$$\begin{aligned} \phi_k &\sim H \\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\ z_i|\pi &\sim \text{Discrete}(\pi) \\ x_i|\phi_{z_i} &\sim F(\cdot|\phi_{z_i}) \end{aligned}$$



- Model selection/averaging over:
  - Hyperparameters in H.
  - Dirichlet parameter α.
  - Number of components K.
- Determining K hardest.

Infinite Mixture Models

- Imagine that $K \gg 0$ is really large.
- If parameters ϕ_k and mixing proportions π integrated out, the number of latent variables left does not grow with K—no overfitting.
- At most n components will be associated with data, aka "active".
- Usually, the number of active components is much less than n.
- This gives an infinite mixture model.
- Demo: dpm_demo2d
- *Issue 1: can we take this limit $K \to \infty$?*
- *Issue 2: what is the corresponding limiting model?*



[Rasmussen 2000]

- A Gaussian process (GP) is a distribution over functions

$$f : \mathbb{X} \mapsto \mathbb{R}$$

- Denote $f \sim \text{GP}$ if f is a GP-distributed random function.
- For any finite set of input points $x_1, \ldots, x_n$, we require $(f(x_1), \ldots, f(x_n))$ to be a multivariate Gaussian.

- The GP is parametrized by its mean $m(x)$ and covariance $c(x, y)$ functions:

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} c(x_1, x_1) & \dots & c(x_1, x_n) \\ \vdots & \ddots & \vdots \\ c(x_n, x_1) & \dots & c(x_n, x_n) \end{bmatrix} \right)$$

- The above are finite dimensional marginal distributions of the GP.
- A salient property of these marginal distributions is that they are consistent: integrating out variables preserves the parametric form of the marginal distributions above.

- A sequence of input points $x_1, x_2, x_3, \ldots$ dense in $\mathbb{X}$.
- Draw

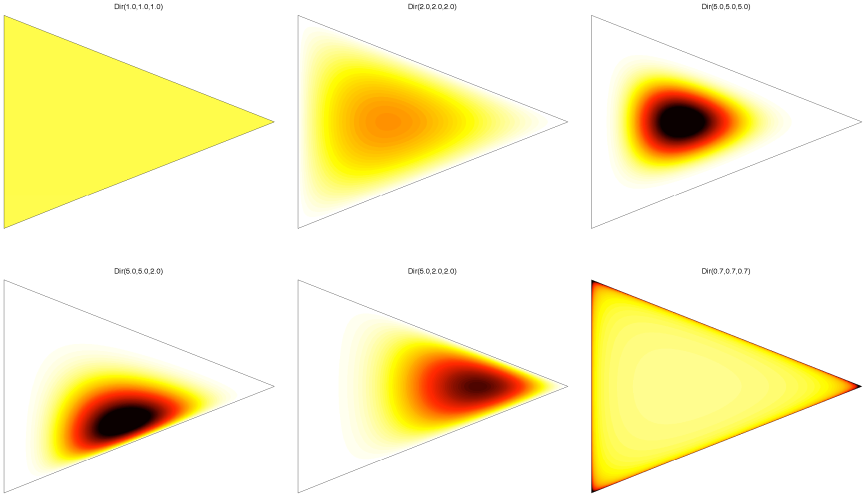
$$\begin{aligned} &f(x_1) \\ &f(x_2) \mid f(x_1) \\ &f(x_3) \mid f(x_1), f(x_2) \\ &\quad\vdots \end{aligned}$$

- Each conditional distribution is Gaussian since $(f(x_1), \ldots, f(x_n))$ is Gaussian.
- Demo: GPgenerate

- A Dirichlet distribution is a distribution over the K-dimensional probability simplex:
$$\Delta_K = \left\{(\pi_1, \ldots, \pi_K) \,:\, \pi_k \geq 0, \textstyle\sum_k \pi_k = 1\right\}$$
- We say $(\pi_1, \ldots, \pi_K)$ is Dirichlet distributed,
$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$
with parameters $(\alpha_1, \ldots, \alpha_K)$, if
$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{n} \pi_k^{\alpha_k - 1}$$

Dirichlet Processes

Examples of Dirichlet distributions.



Dirichlet Processes

Agglomerative property of Dirichlet distributions.

- Combining entries of probability vectors preserves Dirichlet property, for example:

$$\begin{aligned} & (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ \Rightarrow \qquad & (\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K) \end{aligned}$$

- Generally, if $(I_1, \ldots, I_j)$ is a partition of $(1, \ldots, n)$:

$$\left(\sum_{i \in I_1} \pi_i, \ldots, \sum_{i \in I_j} \pi_i \right) \sim \text{Dirichlet} \left(\sum_{i \in I_1} \alpha_i, \ldots, \sum_{i \in I_j} \alpha_i \right)$$

- The converse of the agglomerative property is also true, for example if:

$$\begin{aligned}(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ (\tau_1, \tau_2) &\sim \text{Dirichlet}(\alpha_1\beta_1, \alpha_1\beta_2)\end{aligned}$$

with $\beta_1 + \beta_2 = 1$,

$$\Rightarrow \quad (\pi_1\tau_1, \pi_1\tau_2, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1\beta_1, \alpha_2\beta_2, \alpha_2, \ldots, \alpha_K)$$

- A Dirichlet process (DP) is an “infinitely decimated” Dirichlet distribution:

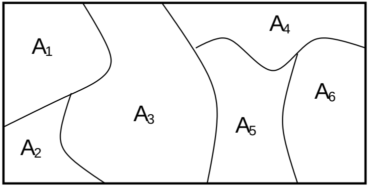
$$
\begin{aligned}
1 &\sim \mathrm{Dirichlet}(\alpha) \\
(\pi_1, \pi_2) &\sim \mathrm{Dirichlet}(\alpha/2, \alpha/2) & \pi_1 + \pi_2 = 1 \\
(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) &\sim \mathrm{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4) & \pi_{i1} + \pi_{i2} = \pi_i \\
&\vdots
\end{aligned}
$$

- Each decimation step involves drawing from a Beta distribution (a Dirichlet with 2 components) and multiplying into the relevant entry.
- Demo: DPgenerate

Dirichlet Processes

A Proper but Non-Constructive Definition

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A Dirichlet Process (DP) is a distribution over probability measures.
- Denote $G \sim \text{DP}$ if G is a DP-distributed random probability measure.
- For any finite set of partitions $A_1 \dot\cup \ldots \dot\cup A_K = \mathbb{X}$, we require $(G(A_1), \ldots, G(A_K))$ to be Dirichlet distributed.



- A DP has two parameters:
  - Base distribution H, which is like the *mean* of the DP.
  - Strength parameter α, which is like an *inverse-variance* of the DP.
- We write:

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

if for any partition $(A_1, \ldots, A_K)$ of $\mathbb{X}$:

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

- A DP has two parameters:
  - Base distribution H, which is like the *mean* of the DP.
  - Strength parameter α, which is like an *inverse-variance* of the DP.
- We write:

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

- The first two cumulants of the DP:

$$\text{Expectation:} \qquad \mathbb{E}[G(A)] = H(A)$$

$$\text{Variance:} \qquad \mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

  where A is any measurable subset of $\mathbb{X}$.

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: [Ferguson 1973].
- de Finetti's Theorem: Blackwell-MacQueen urn scheme, Chinese restaurant process, [Blackwell and MacQueen 1973, Aldous 1985].
- Stick-breaking Construction: [Sethuraman 1994].
- Gamma Process: [Ferguson 1973].

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: [Ferguson 1973].
- de Finetti's Theorem: Blackwell-MacQueen urn scheme, Chinese restaurant process, [Blackwell and MacQueen 1973, Aldous 1985].
- Stick-breaking Construction: [Sethuraman 1994].
- Gamma Process: [Ferguson 1973].

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: [Ferguson 1973].
- de Finetti's Theorem: Blackwell-MacQueen urn scheme, Chinese restaurant process, [Blackwell and MacQueen 1973, Aldous 1985].
- Stick-breaking Construction: [Sethuraman 1994].
- Gamma Process: [Ferguson 1973].

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: [Ferguson 1973].
- de Finetti's Theorem: Blackwell-MacQueen urn scheme, Chinese restaurant process, [Blackwell and MacQueen 1973, Aldous 1985].
- Stick-breaking Construction: [Sethuraman 1994].
- Gamma Process: [Ferguson 1973].

- A probability measure is a function from subsets of a space $\mathbb{X}$ to $[0, 1]$ satisfying certain properties.
- A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- How do we know that such an object exists?!?
- Kolmogorov Consistency Theorem: [Ferguson 1973].
- de Finetti's Theorem: Blackwell-MacQueen urn scheme, Chinese restaurant process, [Blackwell and MacQueen 1973, Aldous 1985].
- Stick-breaking Construction: [Sethuraman 1994].
- Gamma Process: [Ferguson 1973].

Outline

- Suppose $G \sim \mathrm{DP}(\alpha, H)$. G is a (random) probability measure over $\mathbb{X}$. We can treat it as a distribution over $\mathbb{X}$. Let

$$\theta_1, \ldots, \theta_n \sim G$$

  be a random variable with distribution G.

- We saw in the demo that draws from a Dirichlet process seem to be discrete distributions. If so, then:

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

  and there is positive probability that θ_i's can take on the same value θ_k^* for some k, i.e. the θ_i's cluster together.

- In this section we are concerned with representations of Dirichlet processes based upon both the clustering property and the sum of point masses.

- Suppose $G \sim \text{DP}(\alpha, H)$. G is a (random) probability measure over $\mathbb{X}$. We can treat it as a distribution over $\mathbb{X}$. Let

$$\theta_1, \ldots, \theta_n \sim G$$

  be a random variable with distribution G.

- We saw in the demo that draws from a Dirichlet process seem to be discrete distributions. If so, then:

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

  and there is positive probability that θ_i's can take on the same value θ_k^* for some k, i.e. the θ_i's cluster together.

- In this section we are concerned with representations of Dirichlet processes based upon both the clustering property and the sum of point masses.

- Suppose $G \sim \mathrm{DP}(\alpha, H)$. G is a (random) probability measure over $\mathbb{X}$. We can treat it as a distribution over $\mathbb{X}$. Let

$$\theta_1, \ldots, \theta_n \sim G$$

be a random variable with distribution G.

- We saw in the demo that draws from a Dirichlet process seem to be discrete distributions. If so, then:

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

and there is positive probability that θ_i's can take on the same value θ_k^* for some k, i.e. the θ_i's cluster together.

- In this section we are concerned with representations of Dirichlet processes based upon both the clustering property and the sum of point masses.

- Suppose G is Dirichlet process distributed:

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

- G is a (random) probability measure over $\mathbb{X}$. We can treat it as a distribution over $\mathbb{X}$. Let

$$\theta \sim G$$

  be a random variable with distribution G.

- We are interested in:

$$p(\theta) = \int p(\theta|G)p(G)\,dG$$

$$p(G|\theta) = \frac{p(\theta|G)p(G)}{p(\theta)}$$

- Consider:

$$(\pi_1, \dots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$$
$$z|(\pi_1, \dots, \pi_K) \sim \text{Discrete}(\pi_1, \dots, \pi_K)$$

z is a multinomial variate, taking on value $i \in \{1, \dots, n\}$ with probability π_i.

- Then:

$$z \sim \text{Discrete}\left(\frac{\alpha_1}{\sum_i \alpha_i}, \dots, \frac{\alpha_K}{\sum_i \alpha_i}\right)$$
$$(\pi_1, \dots, \pi_K)|z \sim \text{Dirichlet}(\alpha_1 + \delta_1(z), \dots, \alpha_K + \delta_K(z))$$

where $\delta_i(z) = 1$ if z takes on value i, 0 otherwise.

- Converse also true.

Posterior Dirichlet Processes

- Fix a partition $(A_1, \ldots, A_K)$ of $\mathbb{X}$. Then

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$
$$P(\theta \in A_i | G) = G(A_i)$$

- Using Dirichlet-multinomial conjugacy,

$$P(\theta \in A_i) = H(A_i)$$
$$(G(A_1), \ldots, G(A_K))|\theta \sim \text{Dirichlet}(\alpha H(A_1)+\delta_\theta(A_1), \ldots, \alpha H(A_K)+\delta_\theta(A_K))$$

- The above is true for every finite partition of $\mathbb{X}$. In particular, taking a really fine partition,

$$p(\theta)d\theta = H(d\theta)$$

- Also, the posterior $G|\theta$ is also a Dirichlet process:

$$G|\theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$$

Posterior Dirichlet Processes

- Fix a partition $(A_1, \dots, A_K)$ of $\mathbb{X}$. Then

$$(G(A_1), \dots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \dots, \alpha H(A_K))$$
$$P(\theta \in A_i | G) = G(A_i)$$

- Using Dirichlet-multinomial conjugacy,

$$P(\theta \in A_i) = H(A_i)$$
$$(G(A_1), \dots, G(A_K))|\theta \sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \dots, \alpha H(A_K) + \delta_\theta(A_K))$$

- The above is true for every finite partition of $\mathbb{X}$. In particular, taking a really fine partition,

$$p(\theta)d\theta = H(d\theta)$$

- Also, the posterior $G|\theta$ is also a Dirichlet process:

$$G|\theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$$

Posterior Dirichlet Processes

- Fix a partition $(A_1, \ldots, A_K)$ of $\mathbb{X}$. Then

$$\begin{aligned}(G(A_1), \ldots, G(A_K)) &\sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)) \\ P(\theta \in A_i | G) &= G(A_i)\end{aligned}$$

- Using Dirichlet-multinomial conjugacy,

$$\begin{aligned}P(\theta \in A_i) &= H(A_i) \\ (G(A_1), \ldots, G(A_K))|\theta &\sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \ldots, \alpha H(A_K) + \delta_\theta(A_K))\end{aligned}$$

- The above is true for every finite partition of $\mathbb{X}$. In particular, taking a really fine partition,

$$p(\theta)d\theta = H(d\theta)$$

- Also, the posterior $G|\theta$ is also a Dirichlet process:

$$G|\theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$$

Posterior Dirichlet Processes

- Fix a partition $(A_1, \ldots, A_K)$ of $\mathbb{X}$. Then

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$
$$P(\theta \in A_i | G) = G(A_i)$$

- Using Dirichlet-multinomial conjugacy,

$$P(\theta \in A_i) = H(A_i)$$
$$(G(A_1), \ldots, G(A_K))|\theta \sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \ldots, \alpha H(A_K) + \delta_\theta(A_K))$$

- The above is true for every finite partition of $\mathbb{X}$. In particular, taking a really fine partition,

$$p(\theta) d\theta = H(d\theta)$$

- Also, the posterior $G|\theta$ is also a Dirichlet process:

$$G|\theta \sim \text{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$$

Posterior Dirichlet Processes

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \theta &\sim H \\ G | \theta &\sim \mathrm{DP}\left(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1}\right) \end{aligned}$$

Blackwell-MacQueen Urn Scheme

- First sample:

$$\begin{aligned} & \theta_1 | G \sim G & & G \sim \mathrm{DP}(\alpha, H) \\ \iff & \theta_1 \sim H & & G|\theta_1 \sim \mathrm{DP}(\alpha + 1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha + 1}) \end{aligned}$$

- Second sample:

$$\begin{aligned} & \theta_2 | \theta_1, G \sim G & & G|\theta_1 \sim \mathrm{DP}(\alpha + 1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha + 1}) \\ \iff & \theta_2 | \theta_1 \sim \tfrac{\alpha H + \delta_{\theta_1}}{\alpha + 1} & & G|\theta_1, \theta_2 \sim \mathrm{DP}(\alpha + 2, \tfrac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha + 2}) \end{aligned}$$

- n^{th} sample

$$\begin{aligned} & \theta_n | \theta_{1:n-1}, G \sim G & & G|\theta_{1:n-1} \sim \mathrm{DP}(\alpha + n - 1, \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}) \\ \iff & \theta_n | \theta_{1:n-1} \sim \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1} & & G|\theta_{1:n} \sim \mathrm{DP}(\alpha + n, \tfrac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}) \end{aligned}$$

Blackwell-MacQueen Urn Scheme

- First sample:

$$\begin{array}{rrr} & \theta_1 | G \sim G & G \sim \mathrm{DP}(\alpha, H) \\ \iff & \theta_1 \sim H & G|\theta_1 \sim \mathrm{DP}(\alpha+1, \frac{\alpha H + \delta_{\theta_1}}{\alpha+1}) \end{array}$$

- Second sample:

$$\begin{array}{rrr} & \theta_2 | \theta_1, G \sim G & G|\theta_1 \sim \mathrm{DP}(\alpha+1, \frac{\alpha H + \delta_{\theta_1}}{\alpha+1}) \\ \iff & \theta_2|\theta_1 \sim \frac{\alpha H + \delta_{\theta_1}}{\alpha+1} & G|\theta_1, \theta_2 \sim \mathrm{DP}(\alpha+2, \frac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha+2}) \end{array}$$

- n^{th} sample

$$\begin{array}{rrr} & \theta_n | \theta_{1:n-1}, G \sim G & G|\theta_{1:n-1} \sim \mathrm{DP}(\alpha+n-1, \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1}) \\ \iff & \theta_n|\theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} & G|\theta_{1:n} \sim \mathrm{DP}(\alpha+n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n}) \end{array}$$

Blackwell-MacQueen Urn Scheme

- First sample:

$$\theta_1|G \sim G \qquad G \sim \mathrm{DP}(\alpha, H)$$

$$\iff \qquad \theta_1 \sim H \qquad G|\theta_1 \sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1})$$

- Second sample:

$$\theta_2|\theta_1, G \sim G \qquad G|\theta_1 \sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1})$$

$$\iff \qquad \theta_2|\theta_1 \sim \tfrac{\alpha H + \delta_{\theta_1}}{\alpha+1} \qquad G|\theta_1, \theta_2 \sim \mathrm{DP}(\alpha+2, \tfrac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha+2})$$

- n^{th} sample

$$\theta_n|\theta_{1:n-1}, G \sim G \qquad G|\theta_{1:n-1} \sim \mathrm{DP}(\alpha+n-1, \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1})$$

$$\iff \qquad \theta_n|\theta_{1:n-1} \sim \tfrac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} \qquad G|\theta_{1:n} \sim \mathrm{DP}(\alpha+n, \tfrac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha+n})$$

Blackwell-MacQueen Urn Scheme

- Blackwell-MacQueen urn scheme produces a sequence $\theta_1, \theta_2, \ldots$ with the following conditionals:

$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- Picking balls of different colors from an urn:
  - Start with no balls in the urn.
  - with probability $\propto \alpha$, draw $\theta_n \sim H$, and add a ball of that color into the urn.
  - With probability $\propto n - 1$, pick a ball at random from the urn, record θ_n to be its color, return the ball into the urn and place a second ball of same color into urn.
- Blackwell-MacQueen urn scheme is like a "representer" for the DP—a finite projection of an infinite object.

Exchangeability and de Finetti's Theorem

- Starting with a DP, we constructed Blackwell-MacQueen urn scheme.
- The reverse is possible using de Finetti's Theorem.
- Since θ_i are iid $\sim G$, their joint distribution is invariant to permutations, thus $\theta_1, \theta_2, \ldots$ are exchangeable.
- Thus a distribution over measures must exist making them iid.
- This is the DP.

Chinese Restaurant Process

- Draw $\theta_1, \ldots, \theta_n$ from a Blackwell-MacQueen urn scheme.
- They take on $K < n$ distinct values, say $\theta_1^*, \ldots, \theta_K^*$.
- This defines a partition of $1, \ldots, n$ into K clusters, such that if i is in cluster k, then $\theta_i = \theta_k^*$.
- Random draws $\theta_1, \ldots, \theta_n$ from a Blackwell-MacQueen urn scheme induces a random partition of $1, \ldots, n$.
- The induced distribution over partitions is a Chinese restaurant process (CRP).

Chinese Restaurant Process

- Generating from the CRP:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
  - Customers $\Leftrightarrow$ integers, tables $\Leftrightarrow$ clusters.
- The CRP exhibits the clustering property of the DP.

Chinese Restaurant Process

- Generating from the CRP:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
  - Customers $\Leftrightarrow$ integers, tables $\Leftrightarrow$ clusters.
- The CRP exhibits the clustering property of the DP.

Chinese Restaurant Process

- To get back from the CRP to Blackwell-MacQueen urn scheme, simply draw

$$\theta_k^* \sim H$$

for $k = 1, \ldots, K$, then for $i = 1, \ldots, n$ set

$$\theta_i = \theta_{z_i}^*$$

where z_i is the table that customer i sat at.

- The CRP teases apart the clustering property of the DP, from the base distribution.

Stick-breaking Construction

- Returning to the posterior process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta &\sim G \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \theta &\sim H \\ G &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_\theta}{\alpha+1}) \end{aligned}$$

- Consider a partition $(\theta, \mathbb{X}\backslash\theta)$ of $\mathbb{X}$. We have:

$$\begin{aligned} (G(\theta), G(\mathbb{X}\backslash\theta)) &\sim \mathrm{Dirichlet}((\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\theta), (\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\mathbb{X}\backslash\theta)) \\ &= \mathrm{Dirichlet}(1, \alpha) \end{aligned}$$

- G has a point mass located at θ:

$$G = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \mathrm{Beta}(1, \alpha)$$

and G' is the (renormalized) probability measure with the point mass removed.

- What is G'?

Stick-breaking Construction

- Returning to the posterior process:

$$G \sim \mathrm{DP}(\alpha, H) \qquad \theta \sim H$$
$$\theta \sim G \qquad \Leftrightarrow \qquad G \sim \mathrm{DP}(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha + 1})$$

- Consider a partition $(\theta, \mathbb{X} \backslash \theta)$ of $\mathbb{X}$. We have:

$$\begin{aligned}(G(\theta), G(\mathbb{X}\backslash\theta)) &\sim \mathrm{Dirichlet}((\alpha+1)\tfrac{\alpha H + \delta_\theta}{\alpha+1}(\theta), (\alpha+1)\tfrac{\alpha H + \delta_\theta}{\alpha+1}(\mathbb{X}\backslash\theta)) \\ &= \mathrm{Dirichlet}(1, \alpha)\end{aligned}$$

- G has a point mass located at θ:

$$G = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \mathrm{Beta}(1, \alpha)$$

and G' is the (renormalized) probability measure with the point mass removed.

- What is G'?

Stick-breaking Construction

- Returning to the posterior process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta &\sim G \end{aligned} \qquad \Leftrightarrow \qquad \begin{aligned} \theta &\sim H \\ G &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_\theta}{\alpha+1}) \end{aligned}$$

- Consider a partition $(\theta, \mathbb{X}\backslash\theta)$ of $\mathbb{X}$. We have:

$$\begin{aligned} (G(\theta), G(\mathbb{X}\backslash\theta)) &\sim \mathrm{Dirichlet}((\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\theta), (\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\mathbb{X}\backslash\theta)) \\ &= \mathrm{Dirichlet}(1, \alpha) \end{aligned}$$

- G has a point mass located at θ:

$$G = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \mathrm{Beta}(1, \alpha)$$

  and G' is the (renormalized) probability measure with the point mass removed.

- What is G'?

Stick-breaking Construction

- Returning to the posterior process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta &\sim G \end{aligned} \qquad \Leftrightarrow \qquad \begin{aligned} \theta &\sim H \\ G &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_\theta}{\alpha+1}) \end{aligned}$$

- Consider a partition $(\theta, \mathbb{X}\backslash\theta)$ of $\mathbb{X}$. We have:

$$\begin{aligned} (G(\theta), G(\mathbb{X}\backslash\theta)) &\sim \mathrm{Dirichlet}((\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\theta), (\alpha+1)\tfrac{\alpha H+\delta_\theta}{\alpha+1}(\mathbb{X}\backslash\theta)) \\ &= \mathrm{Dirichlet}(1, \alpha) \end{aligned}$$

- G has a point mass located at θ:

$$G = \beta\delta_\theta + (1-\beta)G' \qquad \text{with} \qquad \beta \sim \mathrm{Beta}(1, \alpha)$$

  and G' is the (renormalized) probability measure with the point mass removed.

- What is G'?

Stick-breaking Construction

- Currently, we have:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta &\sim G \end{aligned} \qquad \Rightarrow \qquad \begin{aligned} G &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H + \delta_\theta}{\alpha+1}) \\ G &= \beta\delta_\theta + (1-\beta)G' \\ \theta &\sim H \\ \beta &\sim \mathrm{Beta}(1, \alpha) \end{aligned}$$

- Consider a further partition $(\theta, A_1, \ldots, A_K)$ of $\mathbb{X}$:

$$\begin{aligned} &(G(\theta), G(A_1), \ldots, G(A_K)) \\ =&(\beta, (1-\beta)G'(A_1), \ldots, (1-\beta)G'(A_K)) \end{aligned}$$

- The agglomerative/decimative property of Dirichlet implies:

$$(G'(A_1), \ldots, G'(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$
$$G' \sim \mathrm{DP}(\alpha, H)$$

Stick-breaking Construction

- Currently, we have:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta &\sim G \end{aligned} \qquad \Rightarrow \qquad \begin{aligned} G &\sim \mathrm{DP}(\alpha+1, \tfrac{\alpha H+\delta_\theta}{\alpha+1}) \\ G &= \beta\delta_\theta + (1-\beta)G' \\ \theta &\sim H \\ \beta &\sim \mathrm{Beta}(1,\alpha) \end{aligned}$$

- Consider a further partition $(\theta, A_1, \ldots, A_K)$ of $\mathbb{X}$:

$$\begin{aligned} &(G(\theta), G(A_1), \ldots, G(A_K)) \\ =&(\beta, (1-\beta)G'(A_1), \ldots, (1-\beta)G'(A_K)) \\ \sim&\,\mathrm{Dirichlet}(1, \alpha H(A_1), \ldots, \alpha H(A_K)) \end{aligned}$$

- The agglomerative/decimative property of Dirichlet implies:

$$\begin{aligned} (G'(A_1), \ldots, G'(A_K)) &\sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)) \\ G' &\sim \mathrm{DP}(\alpha, H) \end{aligned}$$

Stick-breaking Construction

- Currently, we have:

$$\begin{array}{l} G \sim \mathrm{DP}(\alpha, H) \\ \theta \sim G \end{array} \quad \Rightarrow \quad \begin{array}{l} G \sim \mathrm{DP}(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha+1}) \\ G = \beta\delta_\theta + (1-\beta)G' \\ \theta \sim H \\ \beta \sim \mathrm{Beta}(1, \alpha) \end{array}$$

- Consider a further partition $(\theta, A_1, \ldots, A_K)$ of $\mathbb{X}$:

$$\begin{aligned} &(G(\theta), G(A_1), \ldots, G(A_K)) \\ =&(\beta, (1-\beta)G'(A_1), \ldots, (1-\beta)G'(A_K)) \\ \sim&\,\mathrm{Dirichlet}(1, \alpha H(A_1), \ldots, \alpha H(A_K)) \end{aligned}$$

- The agglomerative/decimative property of Dirichlet implies:

$$\begin{aligned} (G'(A_1), \ldots, G'(A_K)) &\sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)) \\ G' &\sim \mathrm{DP}(\alpha, H) \end{aligned}$$

Stick-breaking Construction

- We have:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H) \\
G &= \beta_1 \delta_{\theta_1^*} + (1-\beta_1) G_1 \\
G &= \beta_1 \delta_{\theta_1^*} + (1-\beta_1)(\beta_2 \delta_{\theta_2^*} + (1-\beta_2) G_2) \\
&\vdots \\
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}
\end{aligned}
$$

where

$$
\pi_k = \beta_k \prod_{i=1}^{k-1} (1-\beta_i) \qquad \beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \theta_k^* \sim H
$$

- This is the stick-breaking construction.
- Demo: SBgenerate

Stick-breaking Construction

- Starting with a DP, we showed that draws from the DP looks like a sum of point masses, with masses drawn from a stick-breaking construction.
- The steps are limited by assumptions of regularity on $\mathbb{X}$ and smoothness on H.
- [Sethuraman 1994] started with the stick-breaking construction, and showed that draws are indeed DP distributed, under very general conditions.

- Posterior Dirichlet process:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta | G &\sim G \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \theta &\sim H \\ G|\theta &\sim \mathrm{DP}\left(\alpha + 1, \tfrac{\alpha H + \delta_\theta}{\alpha+1}\right) \end{aligned}$$

- Blackwell-MacQueen urn scheme:

$$\theta_n | \theta_{1:n-1} \sim \frac{\alpha H + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

- Chinese restaurant process:

$$p(\text{customer } n \text{ sat at table } k | \text{past}) = \begin{cases} \frac{n_k}{n-1+\alpha} & \text{if occupied table} \\ \frac{\alpha}{n-1+\alpha} & \text{if new table} \end{cases}$$

- Stick-breaking construction:

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \qquad \beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \theta_k^* \sim H \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

Outline

Density Estimation

- Recall our approach to density estimation with Dirichlet processes:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H) \\
x_i &\sim G
\end{aligned}
$$

- The above does not work. Why?
- Problem: G is a discrete distribution; in particular it has no density!
- Solution: Convolve the DP with a smooth distribution:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha, H) \\
F_x(\cdot) &= \int F(\cdot|\theta) dG(\theta) \\
x_i &\sim F_x
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \\
F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\
x_i &\sim F_x
\end{aligned}
$$

Density Estimation

- Recall our approach to density estimation with Dirichlet processes:

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

- The above does not work. Why?
- Problem: G is a discrete distribution; in particular it has no density!
- Solution: Convolve the DP with a smooth distribution:

$$G \sim \mathrm{DP}(\alpha, H)$$
$$F_x(\cdot) = \int F(\cdot|\theta)dG(\theta)$$
$$x_i \sim F_x$$

$$\Rightarrow$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$
$$F_x(\cdot) = \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*)$$
$$x_i \sim F_x$$

Density Estimation

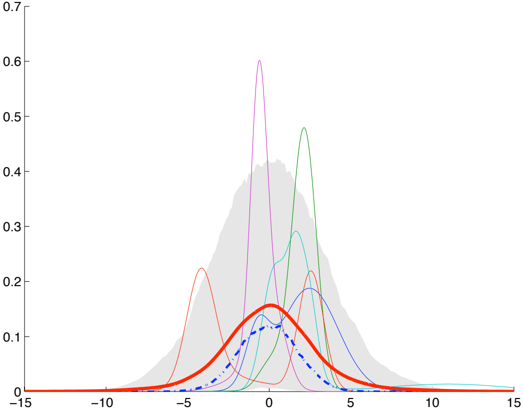
- Recall our approach to density estimation with Dirichlet processes:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ x_i &\sim G \end{aligned}$$

- The above does not work. Why?
- Problem: G is a discrete distribution; in particular it has no density!
- Solution: Convolve the DP with a smooth distribution:

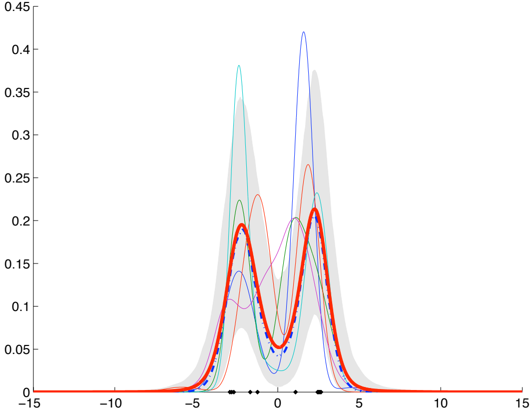
$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ F_x(\cdot) &= \int F(\cdot|\theta)dG(\theta) \\ x_i &\sim F_x \end{aligned} \qquad \Rightarrow \qquad \begin{aligned} G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \\ F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\ x_i &\sim F_x \end{aligned}$$

Density Estimation



$F(\cdot|\mu, \Sigma)$ is Gaussian with mean μ, covariance Σ.
$H(\mu, \Sigma)$ is Gaussian-inverse-Wishart conjugate prior.

Density Estimation



$F(\cdot|\mu,\Sigma)$ is Gaussian with mean μ, covariance Σ.
$H(\mu,\Sigma)$ is Gaussian-inverse-Wishart conjugate prior.

Clustering

- Recall our approach to density estimation:

$$
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \sim \mathrm{DP}(\alpha, H) \\
F_x(\cdot) &= \sum_{k=1}^{\infty} \pi_k F(\cdot|\theta_k^*) \\
x_i &\sim F_x
\end{aligned}
$$

- Above model equivalent to:

$$
\begin{aligned}
z_i &\sim \mathrm{Discrete}(\boldsymbol{\pi}) \\
\theta_i &= \theta_{z_i}^* \\
x_i|z_i &\sim F(\cdot|\theta_i) = F(\cdot|\theta_{z_i}^*)
\end{aligned}
$$

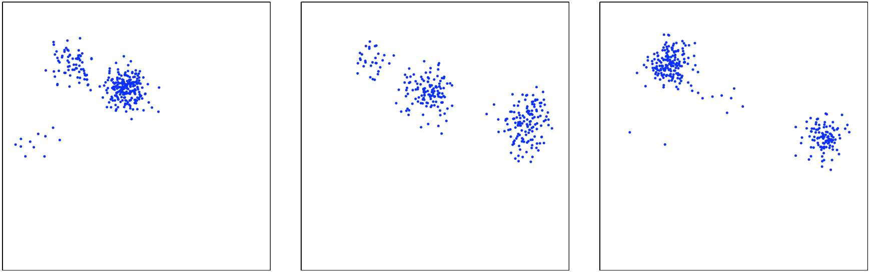
- This is simply a mixture model with an infinite number of components. This is called a DP mixture model.

Clustering

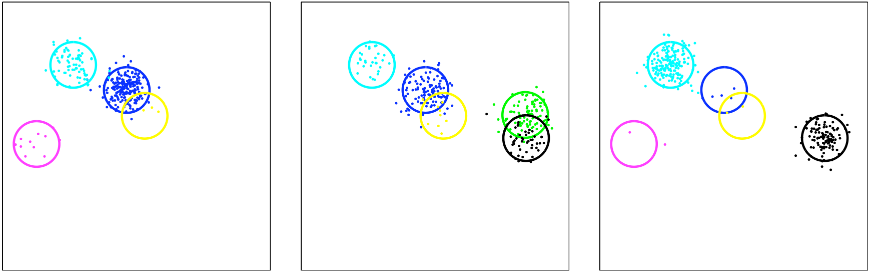
- DP mixture models are used in a variety of clustering applications, where the number of clusters is not known a priori.
- They are also used in applications in which we believe the number of clusters grows without bound as the amount of data grows.
- DPs have also found uses in applications beyond clustering, where the number of latent objects is not known or unbounded.
  - Nonparametric probabilistic context free grammars.
  - Visual scene analysis.
  - Infinite hidden Markov models/trees.
  - Haplotype inference.
  - ...
- In many such applications it is important to be able to model the same set of objects in different contexts.
- This corresponds to the problem of grouped clustering and can be tackled using hierarchical Dirichlet processes.

[Teh et al. 2006]

Grouped Clustering



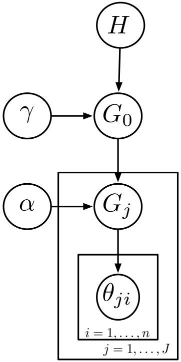
Grouped Clustering



Hierarchical Dirichlet Processes

- Hierarchical Dirichlet process:

$$G_0|\gamma, H \sim \mathrm{DP}(\gamma, H)$$
$$G_j|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$
$$\theta_{ji}|G_j \sim G_j$$

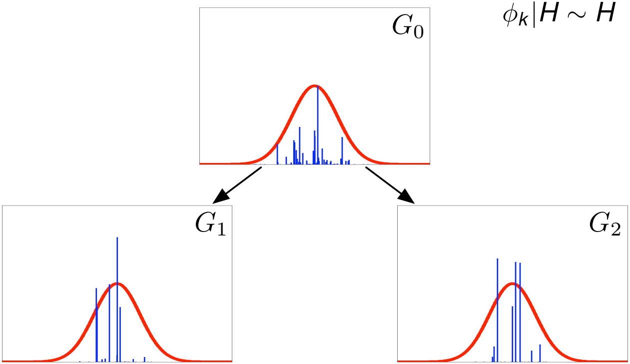


Hierarchical Dirichlet Processes

$$G_0|\gamma, H \sim \mathrm{DP}(\gamma, H) \qquad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \qquad \boldsymbol{\beta}|\gamma \sim \mathrm{Stick}(\gamma)$$

$$G_j|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0) \qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \qquad \boldsymbol{\pi}_j|\alpha, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta})$$

$$\phi_k|H \sim H$$



Summary

- Dirichlet process is “just” a glorified Dirichlet distribution.
- Draws from a DP are probability measures consisting of a weighted sum of point masses.
- Many representations: Blackwell-MacQueen urn scheme, Chinese restaurant process, stick-breaking construction.
- DP mixture models are mixture models with countably infinite number of components.

- I have not delved into:
  - Applications.
  - Generalizations, extensions, other nonparametric processes.
  - Inference: MCMC sampling, variational approximation.
- Also see the tutorial material from Ghaharamani, Jordan and Tresp.

Summary

- Dirichlet process is “just” a glorified Dirichlet distribution.
- Draws from a DP are probability measures consisting of a weighted sum of point masses.
- Many representations: Blackwell-MacQueen urn scheme, Chinese restaurant process, stick-breaking construction.
- DP mixture models are mixture models with countably infinite number of components.

- I have not delved into:
  - Applications.
  - Generalizations, extensions, other nonparametric processes.
  - Inference: MCMC sampling, variational approximation.
- Also see the tutorial material from Ghaharamani, Jordan and Tresp.

Bayesian Nonparametrics

- Parametric models can only capture a bounded amount of information from data, since they have bounded complexity.
- Real data is often complex and the parametric assumption is often wrong.
- Nonparametric models allow relaxation of the parametric assumption, bringing significant flexibility to our models of the world.
- Nonparametric models can also often lead to model selection/averaging behaviours without the cost of actually doing model selection/averaging.
- Nonparametric models are gaining popularity, spurred by growth in computational resources and inference algorithms.
- In addition to DPs, HDPs and their generalizations, other nonparametric models include Indian buffet processes, beta processes, tree processes...

Bayesian Nonparametrics

- Parametric models can only capture a bounded amount of information from data, since they have bounded complexity.
- Real data is often complex and the parametric assumption is often wrong.
- Nonparametric models allow relaxation of the parametric assumption, bringing significant flexibility to our models of the world.
- Nonparametric models can also often lead to model selection/averaging behaviours without the cost of actually doing model selection/averaging.
- Nonparametric models are gaining popularity, spurred by growth in computational resources and inference algorithms.
- In addition to DPs, HDPs and their generalizations, other nonparametric models include Indian buffet processes, beta processes, tree processes...

Bayesian Nonparametrics

- Parametric models can only capture a bounded amount of information from data, since they have bounded complexity.
- Real data is often complex and the parametric assumption is often wrong.
- Nonparametric models allow relaxation of the parametric assumption, bringing significant flexibility to our models of the world.
- Nonparametric models can also often lead to model selection/averaging behaviours without the cost of actually doing model selection/averaging.
- Nonparametric models are gaining popularity, spurred by growth in computational resources and inference algorithms.
- In addition to DPs, HDPs and their generalizations, other nonparametric models include Indian buffet processes, beta processes, tree processes...

Bayesian Nonparametrics

- Parametric models can only capture a bounded amount of information from data, since they have bounded complexity.
- Real data is often complex and the parametric assumption is often wrong.
- Nonparametric models allow relaxation of the parametric assumption, bringing significant flexibility to our models of the world.
- Nonparametric models can also often lead to model selection/averaging behaviours without the cost of actually doing model selection/averaging.
- Nonparametric models are gaining popularity, spurred by growth in computational resources and inference algorithms.
- In addition to DPs, HDPs and their generalizations, other nonparametric models include Indian buffet processes, beta processes, tree processes...

Bayesian Nonparametrics

- Parametric models can only capture a bounded amount of information from data, since they have bounded complexity.
- Real data is often complex and the parametric assumption is often wrong.
- Nonparametric models allow relaxation of the parametric assumption, bringing significant flexibility to our models of the world.
- Nonparametric models can also often lead to model selection/averaging behaviours without the cost of actually doing model selection/averaging.
- Nonparametric models are gaining popularity, spurred by growth in computational resources and inference algorithms.
- In addition to DPs, HDPs and their generalizations, other nonparametric models include Indian buffet processes, beta processes, tree processes...

Outline

Exploring the Dirichlet Process

- Before using DPs, it is important to understand its properties, so that we understand what prior assumptions we are imposing on our models.
- In this practical course we shall work towards implementing a DP mixture model to cluster NIPS papers, thus the relevant properties are the clustering properties of the DP.
- Consider the Chinese restaurant process representation of DPs:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
- How does number of clusters K scale as a function of α and of n (on average)?
- How does the number n_k of customers sitting around table k depend on k and n (on average)?

Exploring the Dirichlet Process

- Before using DPs, it is important to understand its properties, so that we understand what prior assumptions we are imposing on our models.
- In this practical course we shall work towards implementing a DP mixture model to cluster NIPS papers, thus the relevant properties are the clustering properties of the DP.
- Consider the Chinese restaurant process representation of DPs:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
- How does number of clusters K scale as a function of α and of n (on average)?
- How does the number n_k of customers sitting around table k depend on k and n (on average)?

Exploring the Dirichlet Process

- Before using DPs, it is important to understand its properties, so that we understand what prior assumptions we are imposing on our models.
- In this practical course we shall work towards implementing a DP mixture model to cluster NIPS papers, thus the relevant properties are the clustering properties of the DP.
- Consider the Chinese restaurant process representation of DPs:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
- How does number of clusters K scale as a function of α and of n (on average)?
- How does the number n_k of customers sitting around table k depend on k and n (on average)?

Exploring the Dirichlet Process

- Before using DPs, it is important to understand its properties, so that we understand what prior assumptions we are imposing on our models.
- In this practical course we shall work towards implementing a DP mixture model to cluster NIPS papers, thus the relevant properties are the clustering properties of the DP.
- Consider the Chinese restaurant process representation of DPs:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k}{\alpha+n-1}$ where n_k is the number of customers at table k.
    - A new table $K+1$ with probability $\frac{\alpha}{\alpha+n-1}$.
- How does number of clusters K scale as a function of α and of n (on average)?
- How does the number n_k of customers sitting around table k depend on k and n (on average)?

Exploring the Pitman-Yor Process

- Sometimes the assumptions embedded in using DPs to model data are inappropriate.
- The Pitman-Yor process is a generalization of the DP that often has more appropriate properties.
- It has two parameters: d and α with $0 \leq d < 1$ and $\alpha > -d$.

  When $d = 0$ the Pitman-Yor process reduces to a DP.
- It also has a Chinese restaurant process representation:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k - d}{\alpha + n - 1}$ where n_k is the number of customers at table k.
    - A new table $K + 1$ with probability $\frac{\alpha + Kd}{\alpha + n - 1}$.
- How does K scale as a function of n, α and d (on average)?

Exploring the Pitman-Yor Process

- Sometimes the assumptions embedded in using DPs to model data are inappropriate.
- The Pitman-Yor process is a generalization of the DP that often has more appropriate properties.
- It has two parameters: d and α with $0 \leq d < 1$ and $\alpha > -d$.
  When $d = 0$ the Pitman-Yor process reduces to a DP.
- It also has a Chinese restaurant process representation:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k - d}{\alpha + n - 1}$ where n_k is the number of customers at table k.
    - A new table $K + 1$ with probability $\frac{\alpha + Kd}{\alpha + n - 1}$.
- How does K scale as a function of n, α and d (on average)?

Exploring the Pitman-Yor Process

- Sometimes the assumptions embedded in using DPs to model data are inappropriate.
- The Pitman-Yor process is a generalization of the DP that often has more appropriate properties.
- It has two parameters: d and α with $0 \leq d < 1$ and $\alpha > -d$.
  When $d = 0$ the Pitman-Yor process reduces to a DP.
- It also has a Chinese restaurant process representation:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k - d}{\alpha + n - 1}$ where n_k is the number of customers at table k.
    - A new table $K + 1$ with probability $\frac{\alpha + Kd}{\alpha + n - 1}$.
- How does K scale as a function of n, α and d (on average)?

Exploring the Pitman-Yor Process

- Sometimes the assumptions embedded in using DPs to model data are inappropriate.
- The Pitman-Yor process is a generalization of the DP that often has more appropriate properties.
- It has two parameters: d and α with $0 \leq d < 1$ and $\alpha > -d$.
  When $d = 0$ the Pitman-Yor process reduces to a DP.
- It also has a Chinese restaurant process representation:
  - First customer sits at the first table.
  - Customer n sits at:
    - Table k with probability $\frac{n_k - d}{\alpha + n - 1}$ where n_k is the number of customers at table k.
    - A new table $K + 1$ with probability $\frac{\alpha + Kd}{\alpha + n - 1}$.
- How does K scale as a function of n, α and d (on average)?

Dirichlet Process Mixture Models

- We model a data set $x_1, \ldots, x_n$ using the following model:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta_i | G &\sim G \\ x_i | \theta_i &\sim F(\cdot | \theta_i) \qquad\qquad \text{for } i = 1, \ldots, n \end{aligned}$$

- Each θ_i is a latent parameter modelling x_i, while G is the unknown distribution over parameters modelled using a DP.
- This is the basic DP mixture model.
- Implement a DP mixture model.

Dirichlet Process Mixture Models

- We model a data set $x_1, \ldots, x_n$ using the following model:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta_i | G &\sim G \\ x_i | \theta_i &\sim F(\cdot|\theta_i) \qquad\qquad \text{for } i = 1, \ldots, n \end{aligned}$$

- Each θ_i is a latent parameter modelling x_i, while G is the unknown distribution over parameters modelled using a DP.
- This is the basic DP mixture model.
- Implement a DP mixture model.

- Different representations lead to different inference algorithms for DP mixture models.
- The most common are based on the Chinese restaurant process and on the stick-breaking construction.
- Here we shall work with the Chinese restaurant process representation, which, incidentally, can also be derived as the infinite limit of finite mixture models.
- A finite mixture model is defined as follows:

$$
\begin{aligned}
\theta_k^* &\sim H && \text{for } k = 1, \ldots, K \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\
x_i|\theta_{z_i}^* &\sim F(\cdot|\theta_{z_i}^*)
\end{aligned}
$$

- Different representations lead to different inference algorithms for DP mixture models.
- The most common are based on the Chinese restaurant process and on the stick-breaking construction.
- Here we shall work with the Chinese restaurant process representation, which, incidentally, can also be derived as the infinite limit of finite mixture models.
- A finite mixture model is defined as follows:

$$\begin{aligned} \theta_k^* &\sim H & \text{for } k = 1, \ldots, K \\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) & \\ z_i|\pi &\sim \text{Discrete}(\pi) & \text{for } i = 1, \ldots, n \\ x_i|\theta_{z_i}^* &\sim F(\cdot|\theta_{z_i}^*) & \end{aligned}$$

- Different representations lead to different inference algorithms for DP mixture models.
- The most common are based on the Chinese restaurant process and on the stick-breaking construction.
- Here we shall work with the Chinese restaurant process representation, which, incidentally, can also be derived as the infinite limit of finite mixture models.
- A finite mixture model is defined as follows:

$$\begin{aligned} \theta_k^* &\sim H && \text{for } k = 1, \ldots, K \\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) && \\ z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\ x_i|\theta_{z_i}^* &\sim F(\cdot|\theta_{z_i}^*) && \end{aligned}$$

- Different representations lead to different inference algorithms for DP mixture models.
- The most common are based on the Chinese restaurant process and on the stick-breaking construction.
- Here we shall work with the Chinese restaurant process representation, which, incidentally, can also be derived as the infinite limit of finite mixture models.
- A finite mixture model is defined as follows:

$$\begin{aligned}
\theta_k^* &\sim H && \text{for } k = 1, \ldots, K \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\
x_i|\theta_{z_i}^* &\sim F(\cdot|\theta_{z_i}^*)
\end{aligned}$$

- A finite mixture model is defined as follows:

$$\begin{aligned} \phi_k &\sim H && \text{for } k = 1, \ldots, K \\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\ z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\ x_i|\phi_{z_i} &\sim F(\phi_{z_i}) \end{aligned}$$

- Assuming H is conjugate to $F(\cdot|\theta)$, we can integrate out both π and θ_k^*'s, leaving us with z_i's only.
- The simplest MCMC algorithm is to Gibbs sample z_i's (collapsed Gibbs sampling):

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto p(z_i = k|\mathbf{z}_{\neg i})p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$

$$p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) = \int p(x_i|\theta_k^*)p(\theta_k^*|\{x_j : j \neq i, z_j = k\})\, d\theta_k^*$$

- A finite mixture model is defined as follows:

$$\begin{aligned} \phi_k &\sim H && \text{for } k = 1, \ldots, K \\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) && \\ z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\ x_i|\phi_{z_i} &\sim F(\phi_{z_i}) && \end{aligned}$$

- Assuming H is conjugate to $F(\cdot|\theta)$, we can integrate out both π and θ_k^*'s, leaving us with z_i's only.
- The simplest MCMC algorithm is to Gibbs sample z_i's (collapsed Gibbs sampling):

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto p(z_i = k|\mathbf{z}_{\neg i})p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$

$$p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) = \int p(x_i|\theta_k^*)p(\theta_k^*|\{x_j : j \neq i, z_j = k\})\, d\theta_k^*$$

- A finite mixture model is defined as follows:

$$
\begin{aligned}
\phi_k &\sim H && \text{for } k = 1, \ldots, K \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\
x_i|\phi_{z_i} &\sim F(\phi_{z_i})
\end{aligned}
$$

- Assuming H is conjugate to $F(\cdot|\theta)$, we can integrate out both π and θ_k^*'s, leaving us with z_i's only.
- The simplest MCMC algorithm is to Gibbs sample z_i's (collapsed Gibbs sampling):

$$
\begin{aligned}
p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) &\propto p(z_i = k|\mathbf{z}_{\neg i})p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) \\
p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) &= \int p(x_i|\theta_k^*)p(\theta_k^*|\{x_j : j \neq i, z_j = k\})\, d\theta_k^*
\end{aligned}
$$

Aside: Markov Chain Monte Carlo Sampling

- Markov chain Monte Carlo sampling is a dominant and diverse family of inference algorithms for probabilistic models. Here we are interested in obtaining samples from the posterior:

$$\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \theta^*, \pi|\mathbf{x})\, d\theta^* d\pi$$

- The basic idea is to construct a sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ so that for large enough t, $\mathbf{z}^{(t)}$ will be an (approximate) sample from the posterior $p(\mathbf{z}|\mathbf{x})$.
- Convergence to the posterior is guaranteed, but (most of the time) there is no convergence diagnostics, only heuristics. Won't worry about this.
- Given the previous state $\mathbf{z}^{(t-1)}$, we construct $\mathbf{z}^{(t)}$ by making a small (stochastic) alteration to $\mathbf{z}^{(t-1)}$ so that $\mathbf{z}^{(t)}$ is "closer" to the posterior.
- In Gibbs sampling, this alteration is achieved by taking an entry, say z_i, and sampling it from the conditional:

$$z_i^{(t)} \sim p(z_i|\mathbf{z}_{\neg i}^{(t-1)}, \mathbf{x}) \qquad \mathbf{z}_{\neg i}^{(t)} = \mathbf{z}_{\neg i}^{(t-1)}$$

[Neal 1993]

Aside: Markov Chain Monte Carlo Sampling

- Markov chain Monte Carlo sampling is a dominant and diverse family of inference algorithms for probabilistic models. Here we are interested in obtaining samples from the posterior:

$$\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \theta^*, \pi|\mathbf{x})\, d\theta^* d\pi$$

- The basic idea is to construct a sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ so that for large enough t, $\mathbf{z}^{(t)}$ will be an (approximate) sample from the posterior $p(\mathbf{z}|\mathbf{x})$.
- Convergence to the posterior is guaranteed, but (most of the time) there is no convergence diagnostics, only heuristics. Won't worry about this.
- Given the previous state $\mathbf{z}^{(t-1)}$, we construct $\mathbf{z}^{(t)}$ by making a small (stochastic) alteration to $\mathbf{z}^{(t-1)}$ so that $\mathbf{z}^{(t)}$ is "closer" to the posterior.
- In Gibbs sampling, this alteration is achieved by taking an entry, say z_i, and sampling it from the conditional:

$$z_i^{(t)} \sim p(z_i|\mathbf{z}_{\neg i}^{(t-1)}, \mathbf{x}) \qquad\qquad \mathbf{z}_{\neg i}^{(t)} = \mathbf{z}_{\neg i}^{(t-1)}$$

[Neal 1993]

Aside: Markov Chain Monte Carlo Sampling

- Markov chain Monte Carlo sampling is a dominant and diverse family of inference algorithms for probabilistic models. Here we are interested in obtaining samples from the posterior:

$$\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \theta^*, \pi|\mathbf{x})\, d\theta^*\, d\pi$$

- The basic idea is to construct a sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ so that for large enough t, $\mathbf{z}^{(t)}$ will be an (approximate) sample from the posterior $p(\mathbf{z}|\mathbf{x})$.
- Convergence to the posterior is guaranteed, but (most of the time) there is no convergence diagnostics, only heuristics. Won't worry about this.
- Given the previous state $\mathbf{z}^{(t-1)}$, we construct $\mathbf{z}^{(t)}$ by making a small (stochastic) alteration to $\mathbf{z}^{(t-1)}$ so that $\mathbf{z}^{(t)}$ is "closer" to the posterior.
- In Gibbs sampling, this alteration is achieved by taking an entry, say z_i, and sampling it from the conditional:

$$z_i^{(t)} \sim p(z_i|\mathbf{z}_{\neg i}^{(t-1)}, \mathbf{x}) \qquad\qquad \mathbf{z}_{\neg i}^{(t)} = \mathbf{z}_{\neg i}^{(t-1)}$$

[Neal 1993]

Aside: Markov Chain Monte Carlo Sampling

- Markov chain Monte Carlo sampling is a dominant and diverse family of inference algorithms for probabilistic models. Here we are interested in obtaining samples from the posterior:

$$\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \theta^*, \pi|\mathbf{x})\, d\theta^*\, d\pi$$

- The basic idea is to construct a sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ so that for large enough t, $\mathbf{z}^{(t)}$ will be an (approximate) sample from the posterior $p(\mathbf{z}|\mathbf{x})$.
- Convergence to the posterior is guaranteed, but (most of the time) there is no convergence diagnostics, only heuristics. Won't worry about this.
- Given the previous state $\mathbf{z}^{(t-1)}$, we construct $\mathbf{z}^{(t)}$ by making a small (stochastic) alteration to $\mathbf{z}^{(t-1)}$ so that $\mathbf{z}^{(t)}$ is "closer" to the posterior.
- In Gibbs sampling, this alteration is achieved by taking an entry, say z_i, and sampling it from the conditional:

$$z_i^{(t)} \sim p(z_i|\mathbf{z}_{\neg i}^{(t-1)}, \mathbf{x}) \qquad \mathbf{z}_{\neg i}^{(t)} = \mathbf{z}_{\neg i}^{(t-1)}$$

[Neal 1993]

Aside: Markov Chain Monte Carlo Sampling

- Markov chain Monte Carlo sampling is a dominant and diverse family of inference algorithms for probabilistic models. Here we are interested in obtaining samples from the posterior:

$$\mathbf{z}^{(s)} \sim p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \theta^*, \pi|\mathbf{x})\, d\theta^* d\pi$$

- The basic idea is to construct a sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ so that for large enough t, $\mathbf{z}^{(t)}$ will be an (approximate) sample from the posterior $p(\mathbf{z}|\mathbf{x})$.
- Convergence to the posterior is guaranteed, but (most of the time) there is no convergence diagnostics, only heuristics. Won't worry about this.
- Given the previous state $\mathbf{z}^{(t-1)}$, we construct $\mathbf{z}^{(t)}$ by making a small (stochastic) alteration to $\mathbf{z}^{(t-1)}$ so that $\mathbf{z}^{(t)}$ is "closer" to the posterior.
- In Gibbs sampling, this alteration is achieved by taking an entry, say z_i, and sampling it from the conditional:

$$z_i^{(t)} \sim p(z_i|\mathbf{z}_{\neg i}^{(t-1)}, \mathbf{x}) \qquad\qquad \mathbf{z}_{\neg i}^{(t)} = \mathbf{z}_{\neg i}^{(t-1)}$$

[Neal 1993]

Aside: Exponential Families

- An exponential family of distributions is parametrized as:

$$\begin{aligned} p(x|\theta) &= \exp\left(t(\theta)^\top s(x) - \phi(x) - \psi(\theta)\right) \\ s(x) &= \text{sufficient statistics vector.} \\ t(\theta) &= \text{natural parameter vector.} \\ \psi(\theta) &= \log \sum_{x'} \exp\left(t(\theta)^\top s(x') - \phi(x')\right) \quad \text{(log normalization)} \end{aligned}$$

- The conjugate prior is an exponential family distribution over θ:

$$p(\theta) = \exp\left(t(\theta)^\top \nu - \eta\psi(\theta) - \xi(\nu, \eta)\right)$$

- The posterior given observations $x_1, \ldots, x_n$ is in the same family:

$$p(\theta|\mathbf{x}) = \exp\left(t(\theta)^\top (\nu + \textstyle\sum_i s(x_i)) - (\eta + n)\psi(\theta) - \xi(\nu + \textstyle\sum_i s(x_i), \eta + n)\right)$$

- The marginal probability is:

$$p(\mathbf{x}) = \exp\left(\xi(\nu + \textstyle\sum_i s(x_i), \eta + n) - \xi(\nu, \eta) - \textstyle\sum_i \phi(x_i)\right)$$

Aside: Exponential Families

- An exponential family of distributions is parametrized as:

$$\begin{aligned} p(x|\theta) &= \exp\left(t(\theta)^\top s(x) - \phi(x) - \psi(\theta)\right) \\ s(x) &= \text{sufficient statistics vector.} \\ t(\theta) &= \text{natural parameter vector.} \\ \psi(\theta) &= \log \sum_{x'} \exp\left(t(\theta)^\top s(x') - \phi(x')\right) \quad \text{(log normalization)} \end{aligned}$$

- The conjugate prior is an exponential family distribution over θ:

$$p(\theta) = \exp\left(t(\theta)^\top \nu - \eta\psi(\theta) - \xi(\nu,\eta)\right)$$

- The posterior given observations $x_1, \ldots, x_n$ is in the same family:

$$p(\theta|\mathbf{x}) = \exp\left(t(\theta)^\top \left(\nu + \textstyle\sum_i s(x_i)\right) - (\eta+n)\psi(\theta) - \xi(\nu + \textstyle\sum_i s(x_i), \eta+n)\right)$$

- The marginal probability is:

$$p(\mathbf{x}) = \exp\left(\xi(\nu + \textstyle\sum_i s(x_i), \eta + n) - \xi(\nu,\eta) - \textstyle\sum_i \phi(x_i)\right)$$

Aside: Exponential Families

- An exponential family of distributions is parametrized as:

$$
\begin{aligned}
p(x|\theta) &= \exp\left(t(\theta)^\top s(x) - \phi(x) - \psi(\theta)\right) \\
s(x) &= \text{sufficient statistics vector.} \\
t(\theta) &= \text{natural parameter vector.} \\
\psi(\theta) &= \log \sum_{x'} \exp\left(t(\theta)^\top s(x') - \phi(x')\right) \quad \text{(log normalization)}
\end{aligned}
$$

- The conjugate prior is an exponential family distribution over θ:

$$p(\theta) = \exp\left(t(\theta)^\top \nu - \eta\psi(\theta) - \xi(\nu, \eta)\right)$$

- The posterior given observations $x_1, \ldots, x_n$ is in the same family:

$$p(\theta|\mathbf{x}) = \exp\left(t(\theta)^\top (\nu + \textstyle\sum_i s(x_i)) - (\eta + n)\psi(\theta) - \xi(\nu + \sum_i s(x_i), \eta + n)\right)$$

- The marginal probability is:

$$p(\mathbf{x}) = \exp\left(\xi(\nu + \textstyle\sum_i s(x_i), \eta + n) - \xi(\nu, \eta) - \sum_i \phi(x_i)\right)$$

Aside: Exponential Families

- An exponential family of distributions is parametrized as:

$$
\begin{aligned}
p(x|\theta) &= \exp\left(t(\theta)^\top s(x) - \phi(x) - \psi(\theta)\right) \\
s(x) &= \text{sufficient statistics vector.} \\
t(\theta) &= \text{natural parameter vector.} \\
\psi(\theta) &= \log \sum_{x'} \exp\left(t(\theta)^\top s(x') - \phi(x')\right) \;\; \text{(log normalization)}
\end{aligned}
$$

- The conjugate prior is an exponential family distribution over θ:

$$p(\theta) = \exp\left(t(\theta)^\top \nu - \eta\psi(\theta) - \xi(\nu, \eta)\right)$$

- The posterior given observations $x_1, \ldots, x_n$ is in the same family:

$$p(\theta|\mathbf{x}) = \exp\left(t(\theta)^\top (\nu + \textstyle\sum_i s(x_i)) - (\eta + n)\psi(\theta) - \xi(\nu + \textstyle\sum_i s(x_i), \eta + n)\right)$$

- The marginal probability is:

$$p(\mathbf{x}) = \exp\left(\xi(\nu + \textstyle\sum_i s(x_i), \eta + n) - \xi(\nu, \eta) - \textstyle\sum_i \phi(x_i)\right)$$

- Finite mixture model:

$$\begin{aligned} \phi_k &\sim H & \text{for } k = 1, \ldots, K \\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\ z_i|\pi &\sim \text{Discrete}(\pi) & \text{for } i = 1, \ldots, n \\ x_i|\phi_{z_i} &\sim F(\phi_{z_i}) \end{aligned}$$

- Integrating out both π and θ_k^*'s, the Gibbs sampling conditional distributions for $\mathbf{z}$ are:

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto (n_k^{\neg i} + \alpha/K) p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$

$$\begin{aligned} p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) &= \int p(x_i|\theta_k^*) p(\theta_k^*|\{x_j : j \neq i, z_j = k\})\, d\theta_k^* \\ &= \exp\left(\xi(\nu + s(x_i) + \textstyle\sum_{j\neq i: z_j = k} s(x_j), \eta + 1 + n_k^{\neg i}) \right. \\ &\quad \left. - \xi(\nu + \textstyle\sum_{j\neq i: z_j=k} s(x_j), \eta + n_k^{\neg i}) - \phi(x_i)\right) \end{aligned}$$

- Demo: fm_demo2d

- Finite mixture model:

$$\begin{aligned}
\phi_k &\sim H && \text{for } k = 1, \ldots, K \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\
x_i|\phi_{z_i} &\sim F(\phi_{z_i})
\end{aligned}$$

- Integrating out both π and θ_k^*'s, the Gibbs sampling conditional distributions for $\mathbf{z}$ are:

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto (n_k^{\neg i} + \alpha/K)p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$

$$\begin{aligned}
p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) &= \int p(x_i|\theta_k^*)p(\theta_k^*|\{x_j : j \neq i, z_j = k\})\, d\theta_k^* \\
&= \exp\Big(\xi(\nu + s(x_i) + \textstyle\sum_{j\neq i:z_j=k} s(x_j), \eta + 1 + n_k^{\neg i}) \\
&\qquad - \xi(\nu + \textstyle\sum_{j\neq i:z_j=k} s(x_j), \eta + n_k^{\neg i}) - \phi(x_i)\Big)
\end{aligned}$$

- Demo: fm_demo2d

- Finite mixture model:

$$\begin{aligned}
\phi_k &\sim H && \text{for } k = 1, \ldots, K \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
z_i|\pi &\sim \text{Discrete}(\pi) && \text{for } i = 1, \ldots, n \\
x_i|\phi_{z_i} &\sim F(\phi_{z_i})
\end{aligned}$$

- Integrating out both π and θ_k^*'s, the Gibbs sampling conditional distributions for $\mathbf{z}$ are:

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto (n_k^{\neg i} + \alpha/K)p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$

$$\begin{aligned}
p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) &= \int p(x_i|\theta_k^*)p(\theta_k^*|\{x_j : j \neq i, z_j = k\})\, d\theta_k^* \\
&= \exp\left(\xi(\nu + s(x_i) + \textstyle\sum_{j\neq i:z_j=k} s(x_j), \eta + 1 + n_k^{\neg i}) \right. \\
&\qquad \left. - \xi(\nu + \textstyle\sum_{j\neq i:z_j=k} s(x_j), \eta + n_k^{\neg i}) - \phi(x_i)\right)
\end{aligned}$$

- Demo: fm_demo2d

- Imagine that $K \gg 0$ is really large.
- Only a few components will be "active" (i.e. with $n_k > 0$), while most are "inactive".

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto \begin{cases} (n_k^{\neg i} + \alpha/K)p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) & \text{if } k \text{ active;} \\ (\alpha/K)p(x_i) & \text{if } k \text{ inactive.} \end{cases}$$

$$\begin{aligned} p(z_i = k \text{ active}|\mathbf{z}^{\neg i}, \mathbf{x}) &\propto (n_k^{\neg i} + \alpha/K)p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) \\ &\approx n_k^{\neg i} p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) \\ p(z_i \text{ inactive}|\mathbf{z}^{\neg i}, \mathbf{x}) &\propto (\alpha(K - K_{\text{active}})/K)p(x_i) \\ &\approx \alpha p(x_i) \end{aligned}$$

- This gives an inference algorithm for DP mixture models in Chinese restaurant process representation.

- Imagine that $K \gg 0$ is really large.
- Only a few components will be "active" (i.e. with $n_k > 0$), while most are "inactive".

$$p(z_i = k|\mathbf{z}^{\neg i}, \mathbf{x}) \propto \begin{cases} (n_k^{\neg i} + \alpha/K)p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) & \text{if } k \text{ active;} \\ (\alpha/K)p(x_i) & \text{if } k \text{ inactive.} \end{cases}$$

$$\begin{aligned} p(z_i = k \text{ active}|\mathbf{z}^{\neg i}, \mathbf{x}) &\propto (n_k^{\neg i} + \alpha/K)p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) \\ &\approx n_k^{\neg i} p(x_i|\mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) \\ p(z_i \text{ inactive}|\mathbf{z}^{\neg i}, \mathbf{x}) &\propto (\alpha(K - K_{\text{active}})/K)p(x_i) \\ &\approx \alpha p(x_i) \end{aligned}$$

- This gives an inference algorithm for DP mixture models in Chinese restaurant process representation.

- Rearrange mixture component indices so that $1, \ldots, K_{\text{active}}$ are active, and the rest are inactive.

$$p(z_i = k \le K_{\text{active}} | \mathbf{z}^{\neg i}, \mathbf{x}) \propto n_k^{\neg i} p(x_i | \mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$
$$p(z_i > K_{\text{active}} | \mathbf{z}^{\neg i}, \mathbf{x}) \propto \alpha p(x_i)$$

- If z_i takes on an inactive value, instantiate a new active component, and increment K_{active}.
- If $n_k = 0$ for some k during sampling, delete that active component, and decrement K_{active}.

- Rearrange mixture component indices so that $1, \ldots, K_{\text{active}}$ are active, and the rest are inactive.

$$
\begin{aligned}
p(z_i = k \le K_{\text{active}} | \mathbf{z}^{\neg i}, \mathbf{x}) &\propto n_k^{\neg i} p(x_i | \mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i}) \\
p(z_i > K_{\text{active}} | \mathbf{z}^{\neg i}, \mathbf{x}) &\propto \alpha p(x_i)
\end{aligned}
$$

- If z_i takes on an inactive value, instantiate a new active component, and increment K_{active}.
- If $n_k = 0$ for some k during sampling, delete that active component, and decrement K_{active}.

- Rearrange mixture component indices so that $1, \ldots, K_{\text{active}}$ are active, and the rest are inactive.

$$p(z_i = k \leq K_{\text{active}} | \mathbf{z}^{\neg i}, \mathbf{x}) \propto n_k^{\neg i} p(x_i | \mathbf{z}^{\neg i}, \mathbf{x}_k^{\neg i})$$
$$p(z_i > K_{\text{active}} | \mathbf{z}^{\neg i}, \mathbf{x}) \propto \alpha p(x_i)$$

- If z_i takes on an inactive value, instantiate a new active component, and increment K_{active}.
- If $n_k = 0$ for some k during sampling, delete that active component, and decrement K_{active}.

Clustering NIPS Papers

- I have prepared a small subset of NIPS papers for you to try clustering them.
- We concentrate on a small subset of papers, and a small subset of "informative" words.
- Each paper is represented as a bag-of-words. Paper i is represented by a vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iW})$:

$$x_{iw} = c \qquad \text{if word } w \text{ occurs } c \text{ times in paper } i.$$

- Model papers in cluster k using a Multinomial distribution:

$$p(\mathbf{x}_i|\theta_k^*) = \frac{(\sum_w x_{iw})!}{\prod_w x_{iw}!} \prod_w (\theta_{kw}^*)^{x_{iw}}$$

- The conjugate prior for θ_k^* is a Dirichlet:

$$p(\theta_k^*|b) = \frac{\Gamma(\sum_w b_w)}{\prod_w \Gamma(b_w)} \prod_w (\theta_{kw}^*)^{b_w - 1}$$

Clustering NIPS Papers

- I have prepared a small subset of NIPS papers for you to try clustering them.
- We concentrate on a small subset of papers, and a small subset of "informative" words.
- Each paper is represented as a bag-of-words. Paper i is represented by a vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iW})$:

$$x_{iw} = c \qquad \text{if word } w \text{ occurs } c \text{ times in paper } i.$$

- Model papers in cluster k using a Multinomial distribution:

$$p(\mathbf{x}_i|\theta_k^*) = \frac{(\sum_w x_{iw})!}{\prod_w x_{iw}!} \prod_w (\theta_{kw}^*)^{x_{iw}}$$

- The conjugate prior for θ_k^* is a Dirichlet:

$$p(\theta_k^*|b) = \frac{\Gamma(\sum_w b_w)}{\prod_w \Gamma(b_w)} \prod_w (\theta_{kw}^*)^{b_w - 1}$$

Clustering NIPS Papers

- I have prepared a small subset of NIPS papers for you to try clustering them.
- We concentrate on a small subset of papers, and a small subset of "informative" words.
- Each paper is represented as a bag-of-words. Paper i is represented by a vector $\mathbf{x}_i = (x_{i1}, \dots, x_{iW})$:

$$x_{iw} = c \qquad \text{if word } w \text{ occurs } c \text{ times in paper } i.$$

- Model papers in cluster k using a Multinomial distribution:

$$p(\mathbf{x}_i|\theta_k^*) = \frac{(\sum_w x_{iw})!}{\prod_w x_{iw}!} \prod_w (\theta_{kw}^*)^{x_{iw}}$$

- The conjugate prior for θ_k^* is a Dirichlet:

$$p(\theta_k^*|b) = \frac{\Gamma(\sum_w b_w)}{\prod_w \Gamma(b_w)} \prod_w (\theta_{kw}^*)^{b_w - 1}$$

Clustering NIPS Papers

- I have prepared a small subset of NIPS papers for you to try clustering them.
- We concentrate on a small subset of papers, and a small subset of "informative" words.
- Each paper is represented as a bag-of-words. Paper i is represented by a vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iW})$:

$$x_{iw} = c \qquad \text{if word } w \text{ occurs } c \text{ times in paper } i.$$

- Model papers in cluster k using a Multinomial distribution:

$$p(\mathbf{x}_i|\theta^*_k) = \frac{(\sum_w x_{iw})!}{\prod_w x_{iw}!} \prod_w (\theta^*_{kw})^{x_{iw}}$$

- The conjugate prior for θ^*_k is a Dirichlet:

$$p(\theta^*_k|b) = \frac{\Gamma(\sum_w b_w)}{\prod_w \Gamma(b_w)} \prod_w (\theta^*_{kw})^{b_w - 1}$$

Clustering NIPS Papers

- I have prepared a small subset of NIPS papers for you to try clustering them.
- We concentrate on a small subset of papers, and a small subset of "informative" words.
- Each paper is represented as a bag-of-words. Paper i is represented by a vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iW})$:

$$x_{iw} = c \qquad \text{if word } w \text{ occurs } c \text{ times in paper } i.$$

- Model papers in cluster k using a Multinomial distribution:

$$p(\mathbf{x}_i|\theta_k^*) = \frac{(\sum_w x_{iw})!}{\prod_w x_{iw}!} \prod_w (\theta_{kw}^*)^{x_{iw}}$$

- The conjugate prior for θ_k^* is a Dirichlet:

$$p(\theta_k^*|b) = \frac{\Gamma(\sum_w b_w)}{\prod_w \Gamma(b_w)} \prod_w (\theta_{kw}^*)^{b_w - 1}$$

Clustering NIPS Papers
Specifying the Priors

- We shall use a symmetric Dirichlet prior for the cluster parameters θ. Specifically $b_w = b/W$ for some $b > 0$.
- The model:

$$\begin{aligned} H &= \text{Dirichlet}(b/W, \ldots, b/W) \\ G &\sim \text{DP}(\alpha, H) \\ \theta_i &\sim G \\ \mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i) \end{aligned}$$

- Only two numbers to set: α and b.
- α controls the a priori expected number of clusters.
- b controls the number of words assigned to each cluster.
- What are reasonable values for α and b?

Clustering NIPS Papers

Specifying the Priors

- We shall use a symmetric Dirichlet prior for the cluster parameters θ. Specifically $b_w = b/W$ for some $b > 0$.
- The model:

$$\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}$$

- Only two numbers to set: α and b.
- α controls the a priori expected number of clusters.
- b controls the number of words assigned to each cluster.
- What are reasonable values for α and b?

Clustering NIPS Papers

Specifying the Priors

- We shall use a symmetric Dirichlet prior for the cluster parameters θ. Specifically $b_w = b/W$ for some $b > 0$.
- The model:

$$
\begin{aligned}
H &= \text{Dirichlet}(b/W, \dots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}
$$

- Only two numbers to set: α and b.
- α controls the a priori expected number of clusters.
- b controls the number of words assigned to each cluster.
- What are reasonable values for α and b?

Clustering NIPS Papers

Specifying the Priors

- We shall use a symmetric Dirichlet prior for the cluster parameters θ. Specifically $b_w = b/W$ for some $b > 0$.
- The model:

$$\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}$$

- Only two numbers to set: α and b.
- α controls the a priori expected number of clusters.
- b controls the number of words assigned to each cluster.
- What are reasonable values for α and b?

Clustering NIPS Papers

Sensitivity to Priors

- When building models and making inferences, and one does not "trust" ones prior very much, then it is important to perform sensitivity analysis.
- Sensitivity analysis is about determining how much our inference conclusions depend on the setting of the model priors.
- If our conclusions depend strongly on the priors which we don't trust very much, then we cannot trust our conclusions either.
- If our conclusions do not depend strongly on the priors, then we can more strongly trust our conclusions.
- What part of our model should we worry about?

$$
\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}
$$

Clustering NIPS Papers

Sensitivity to Priors

- When building models and making inferences, and one does not "trust" ones prior very much, then it is important to perform sensitivity analysis.
- Sensitivity analysis is about determining how much our inference conclusions depend on the setting of the model priors.
- If our conclusions depend strongly on the priors which we don't trust very much, then we cannot trust our conclusions either.
- If our conclusions do not depend strongly on the priors, then we can more strongly trust our conclusions.
- What part of our model should we worry about?

$$
\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}
$$

Clustering NIPS Papers

Sensitivity to Priors

- When building models and making inferences, and one does not "trust" ones prior very much, then it is important to perform sensitivity analysis.
- Sensitivity analysis is about determining how much our inference conclusions depend on the setting of the model priors.
- If our conclusions depend strongly on the priors which we don't trust very much, then we cannot trust our conclusions either.
- If our conclusions do not depend strongly on the priors, then we can more strongly trust our conclusions.
- What part of our model should we worry about?

$$
\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}
$$

Clustering NIPS Papers

Sensitivity to Priors

- When building models and making inferences, and one does not "trust" ones prior very much, then it is important to perform sensitivity analysis.
- Sensitivity analysis is about determining how much our inference conclusions depend on the setting of the model priors.
- If our conclusions depend strongly on the priors which we don't trust very much, then we cannot trust our conclusions either.
- If our conclusions do not depend strongly on the priors, then we can more strongly trust our conclusions.
- What part of our model should we worry about?

$$
\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}
$$

- When building models and making inferences, and one does not “trust” ones prior very much, then it is important to perform sensitivity analysis.
- Sensitivity analysis is about determining how much our inference conclusions depend on the setting of the model priors.
- If our conclusions depend strongly on the priors which we don’t trust very much, then we cannot trust our conclusions either.
- If our conclusions do not depend strongly on the priors, then we can more strongly trust our conclusions.
- What part of our model should we worry about?

$$
\begin{aligned}
H &= \text{Dirichlet}(b/W, \ldots, b/W) \\
G &\sim \text{DP}(\alpha, H) \\
\theta_i &\sim G \\
\mathbf{x}_i &\sim \text{Multinomial}(n_i, \theta_i)
\end{aligned}
$$

Summary

- We explored some properties of the Dirichlet process.
- We implemented a Dirichlet process mixture model.
- We applied a Dirichlet process mixture model to clustering NIPS papers.
- We considered ways of specifying the hyperparameters of the model, and explored the sensitivity to these hyperparameters.

- Dirichlet processes are not that mysterious or hard.

Summary

- We explored some properties of the Dirichlet process.
- We implemented a Dirichlet process mixture model.
- We applied a Dirichlet process mixture model to clustering NIPS papers.
- We considered ways of specifying the hyperparameters of the model, and explored the sensitivity to these hyperparameters.

- Dirichlet processes are not that mysterious or hard.

References 0

References I

Aldous, D. (1985).
Exchangeability and related topics.
In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin.

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

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004).
Hierarchical topic models and the nested Chinese restaurant process.
In *Advances in Neural Information Processing Systems*, volume 16.

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
Annals of Statistics, 1(2):209–230.

Finkel, J. R., Grenager, T., and Manning, C. D. (2007).
The infinite tree.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Görür, D. (2007).
Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning.
PhD thesis, Technischen Universität Berlin.

References II

Liang, P., Petrov, S., Jordan, M. I., and Klein, D. (2007).

The infinite PCFG using hierarchical Dirichlet processes.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Neal, R. M. (1993).

Probabilistic inference using Markov chain Monte Carlo methods.

Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Rasmussen, C. E. (2000).

The infinite Gaussian mixture model.

In *Advances in Neural Information Processing Systems*, volume 12.

Rasmussen, C. E. and Ghahramani, Z. (2001).

Occam's razor.

In *Advances in Neural Information Processing Systems*, volume 13.

Sethuraman, J. (1994).

A constructive definition of Dirichlet priors.

Statistica Sinica, 4:639–650.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2007).

Describing visual scenes using transformed objects and parts.

To appear in the International Journal of Computer Vision.

References III

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).

Hierarchical Dirichlet processes.

Journal of the American Statistical Association, 101(476):1566–1581.

Wood, F., Goldwater, S., and Black, M. J. (2006).

A non-parametric Bayesian approach to spike sorting.

In *Proceedings of the IEEE Conference on Engineering in Medicine and Biologicial Systems*, volume 28.