Dirichlet process models

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Outline

Introduction

- Choosing the number of components K in a finite mixture can be tricky.

- A natural Bayesian approach is to put a prior on K.

- What if we don't believe there are finitely many components?

- In this case, it is natural to use an infinite mixture model, i.e., a mixture with infinitely many components: $\sum_{k=1}^{\infty} \pi_k f_{\theta_k}$.

- The most common type of infinite mixture is based on the Dirichlet process.

- The Dirichlet process is an example of a nonparametric Bayesian model.

Introduction: Infinite limit of a finite mixture

- In the following sense, Dirichlet process mixtures are a limiting case of finite mixtures.

- Suppose the prior on mixture weights is

$$\pi \sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K).$$

- As $K \to \infty$, the mixture model $\sum_{k=1}^{K} \pi_k f_{\theta_k}$ converges to a Dirichlet process mixture.

- This helps with intuition and can be useful, but K may need to be quite large for the approximation to be close.

- Below, we will define the DP mixture weights in a different and simpler way, rather than working with this infinite limit.

Introduction: Bayesian nonparametrics (BNP)

- The two main types of nonparametric Bayesian models are:
  1. priors on functions (such as Gaussian processes), and
  2. priors on distributions (such as Dirichlet processes).

- The term "process" signifies that these are stochastic processes, that is, infinite-dimensional random objects.

- Roughly, the term "nonparametrics" refers to highly flexible — usually infinite-dimensional — models.

- BNP started as a Bayesian alternative to frequentist nonparametric statistics.

- Frequentist nonparametric methods fall into a few categories:
  1. flexible estimation of functions (such as kernel regression),
  2. flexible estimation of distributions (such as kernel density estimation), and
  3. "distribution-free" hypothesis testing and estimation.

BNP models have found many applications

- astronomy
- epidemiology
- gene expression profiling
- haplotype inference
- medical image analysis
- survival analysis
- extreme value analysis
- meteorology
- econometrics
- phylogenetics
- species delimitation
- computer vision
- classification
- document modeling
- cognitive science
- natural language processing

Outline

Dirichlet process

- The Dirichlet process (DP) is a distribution on discrete probability measures (Ferguson, 1973).
- The DP is special because it has so many nice properties.
- The DP can be broken down into two parts:
  1. Stick-breaking process: A distribution on the set of infinite sequences $(w_1, w_2, \ldots)$ such that $w_k \geq 0$ and $\sum_{k=1}^{\infty} w_k = 1$.
  2. Base distribution H: The distribution of a sequence of random points $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} H$.
- These are combined to make a random discrete probability distribution
$$\sum_{k=1}^{\infty} w_k \delta_{\theta_k}$$
where δ_θ is the unit point mass at θ.

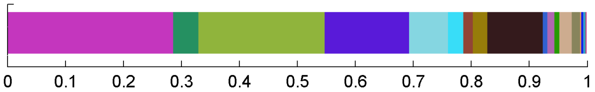
Dirichlet process: Stick-breaking process

- The distribution on weights $w_1, w_2, \ldots$ has an elegant representation due to Sethuraman (1994).

- Definition: Given $\alpha > 0$, if $V_1, V_2, \ldots \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ and

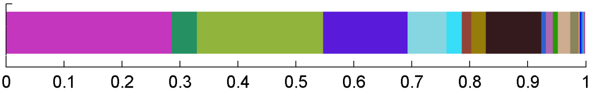
$$W_k = V_k \prod_{i=1}^{k-1} (1 - V_i)$$

for $k = 1, 2, \ldots$, then $(W_1, W_2, \ldots) \sim \text{Stick}(\alpha)$.



Dirichlet process: Stick-breaking process

(Explanation of stick-breaking formula on board)



Dirichlet process: Definition

- Let $\alpha > 0$ and let H be a probability distribution. If

$$\boldsymbol{P} = \sum_{k=1}^{\infty} W_k \delta_{\boldsymbol{\theta}_k}$$

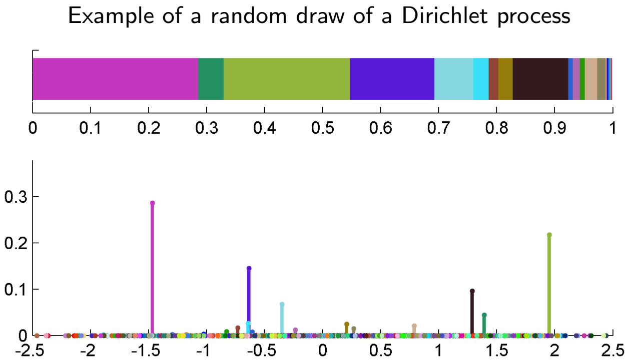
where

$$(W_1, W_2, \ldots) \sim \mathrm{Stick}(\alpha)$$
$$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \overset{\text{iid}}{\sim} H$$

independently, then $\boldsymbol{P} \sim \mathrm{DP}(\alpha, H)$.

- $\boldsymbol{P}$ is a random discrete probability distribution on the same space as H.
- α is called the *concentration parameter*.
- H is called the *base distribution*.

Dirichlet process: Visualization



- Each different vertical line represents a different term $w_k \delta_{\theta_k}$.
- The height is w_k and the location is θ_k.
- In this example, the base distribution H is standard normal.

Dirichlet process: Interpretation of parameters

- The base distribution H is the mean of $\boldsymbol{P}$ in the sense that for any set A,
$$\mathrm{E}\big(\boldsymbol{P}(A)\big) = H(A).$$

- The concentration parameter α controls how close $\boldsymbol{P}$ is to the base distribution H.

- As $\alpha \to \infty$, $\boldsymbol{P}$ converges to H in a certain sense (specifically, in the weak topology).

- Roughly, this is because the weights W_k become smaller and smaller as $\alpha \to \infty$. For instance, $\mathrm{E}(W_1) = 1/(1+\alpha)$.

Dirichlet process: Equivalent definition

- Sethuraman's stick-breaking construction is very nice, but it is not the original definition of the Dirichlet process.
- Ferguson (1973) originally defined the DP as follows.
- Suppose $\boldsymbol{P}$ is a distribution on Θ such that for any partition $\{A_1, \ldots, A_K\}$ of Θ,

  $$(\boldsymbol{P}(A_1), \ldots, \boldsymbol{P}(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)).$$

  Then $\boldsymbol{P} \sim \mathrm{DP}(\alpha, H)$.
- Here, $\mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ is just the finite-dimensional Dirichlet distribution.
- This definition is equivalent to the stick-breaking version. It is more implicit but also has its uses.

Dirichlet process: Posterior distribution

- The posterior of the DP has a simple closed-form expression.

- Consider the following model:

$$\boldsymbol{P} \sim \mathrm{DP}(\alpha, H),$$
$$X_1, \ldots, X_n \mid \boldsymbol{P} = P \overset{\text{iid}}{\sim} P.$$

- Then the posterior on P is

$$\boldsymbol{P}|x_{1:n} \sim \mathrm{DP}(\alpha', H')$$

where $\alpha' = \alpha + n$ and

$$H' = \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n}\Big(\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}\Big).$$

- Thus, we can interpret α as the prior "sample size" and H as a prior guess at the true distibution of the data.

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/ARjuSDTLSZ5HhZeQA

(Three people per room, randomly assigned. 15 minutes.)

Outline

Dirichlet process mixtures

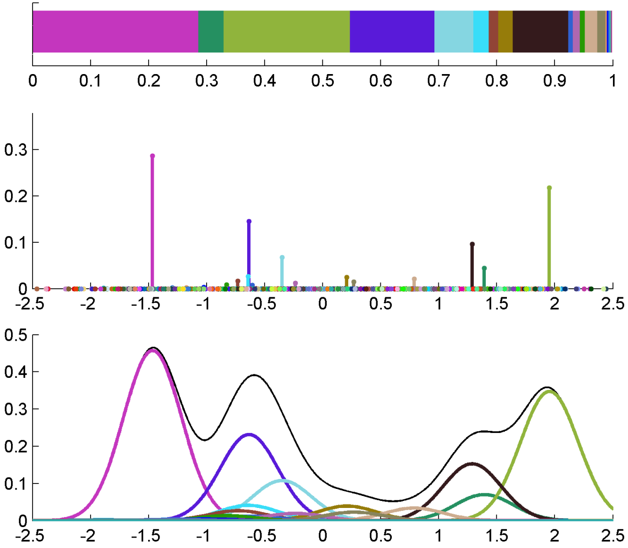
- The DP can be very useful as a prior on distributions.
- However, the fact that $\boldsymbol{P}$ is a discrete distribution has some big limitations in practice.
- Consequently, it is more common to use Dirichlet process mixtures (DPMs).
- In a DPM, the W's and θ's are used as mixture weights and component parameters in a mixture distribution.
- For instance, if $W \sim \mathrm{Stick}(\alpha)$ and $\boldsymbol{\theta}_k := (\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) \overset{\mathrm{iid}}{\sim} H$ then

$$\sum_{k=1}^{\infty} W_k \, \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$$

is a Dirichlet process mixture of Gaussians.

Dirichlet process mixture: Visualization

Example of a random draw of a Dirichlet process mixture of Gaussians



Dirichlet process mixtures: Definition

- More generally, suppose $(f_\theta : \theta \in \Theta)$ is a parametrized family of distributions and H is a distribution on Θ.

- Definition: If $W \sim \mathrm{Stick}(\alpha)$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \overset{\text{iid}}{\sim} H$ then
$$\sum_{k=1}^{\infty} W_k f_{\boldsymbol{\theta}_k}$$
is a *Dirichlet process mixture* (DPM).

- Here, each f_{θ_k} is referred to as a component distribution, and θ_k is the corresponding component parameter.

- In measure-theoretic notation,
$$\sum_{k=1}^{\infty} W_k f_{\boldsymbol{\theta}_k} = \int f_\theta \, d\boldsymbol{P}(\theta)$$
where $\boldsymbol{P} \sim \mathrm{DP}(\alpha, H)$.

Outline

Dirichlet process mixtures: Partition distribution

- The DP induces a distribution on partitions that is very useful for posterior computation in DPMs.

- Any variables $z_1, \ldots, z_n$ induce a partition of $\{1, \ldots, n\}$ such that i and j are in the same part (or "block") if and only if $z_i = z_j$.

- For instance, if $z_{1:6} = (3, 2, 7, 3, 3, 7)$ then the induced partition of $\{1, \ldots, 6\}$ is

$$\mathcal{C} = \mathcal{C}(z) = \Big\{\{1, 4, 5\},\ \{2\},\ \{3, 6\}\Big\}.$$

Restaurant process / Urn scheme

- The DP partition distribution can be described by a sequential sampling scheme.

- This is referred to as the *Chinese restaurant process* (CRP) or *Pólya urn scheme*.

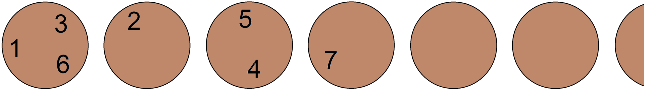
Chinese restaurant process

- The first customer is seated at a table: Initialize $\mathcal{C} = \big\{\{1\}\big\}$.
- For $i = 1, \ldots, n$, the ith customer sits ...

  at table $c \in \mathcal{C}$ with probability $\propto |c|$,

  or at a new table with probability $\propto \alpha$.

  With each new customer, $\mathcal{C}$ is updated to reflect which table they sit at.



Dirichlet process mixtures: Partition distribution

- The DP induces a distribution on partitions as follows.
- Suppose

$$
\begin{aligned}
W &\sim \mathrm{Stick}(\alpha), \\
Z_1, \ldots, Z_n \mid W &\overset{\text{iid}}{\sim} \mathrm{Categorical}(W),
\end{aligned}
$$

  and let $\mathcal{C}$ be the partition of $\{1, \ldots, n\}$ induced by $Z_1, \ldots, Z_n$.

- Integrating out W and $Z_{1:n}$, it turns out that $\mathcal{C}$ has p.m.f.

$$p(\mathcal{C}|\alpha) = \frac{\alpha^{|\mathcal{C}|}\Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{c \in \mathcal{C}} \Gamma\big(|c|\big).$$

- Here, $|\mathcal{C}| =$ number of parts in the partition, $|c| =$ size of part $c \in \mathcal{C}$, and $\Gamma(\cdot)$ is the gamma function.

Dirichlet process mixtures: Partition-based formulation

- A natural way to write a DPM model on data $x_1, \ldots, x_n$ is

$$\begin{aligned}
&W \sim \text{Stick}(\alpha),\\
&Z_1, \ldots, Z_n \mid W \overset{\text{iid}}{\sim} \text{Categorical}(W),\\
&\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \overset{\text{iid}}{\sim} H,\\
&X_i \mid z, \theta \sim f_{\theta_{z_i}} \text{ for } i = 1, \ldots, n.
\end{aligned}$$

- However, for posterior computation, the following equivalent partition-based model is convenient:

$$\begin{aligned}
&\mathcal{C} \sim p(\mathcal{C}|\alpha)\\
&\boldsymbol{\theta}_c \overset{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C},\\
&X_i \mid \mathcal{C}, \theta \sim f_{\theta_c} \text{ for } i \in c,\ c \in \mathcal{C}.
\end{aligned}$$

- $\theta_c \in \Theta$ is the component parameter for the points i in part c.

Individual activity: Quick check

Answer these questions individually (2 minutes):
https://forms.gle/F7h6852eVUooP1xJ9

Outline

Dirichlet process mixtures: Gibbs sampler (1/2)

- The partition-based formulation of the DPM leads to a nice Gibbs sampler algorithm.

- For $c \subseteq \{1, \ldots, n\}$, define

$$m(x_c) := \int \Big(\prod_{i \in c} f_\theta(x_i) \Big) h(\theta) d\theta$$

where $h(\theta)$ is the density of H.

- $m(x_c)$ can be computed analytically when H is a conjugate prior for f_θ.

- For the non-conjugate case, there are also clever MCMC algorithms (Neal, 2000).

Dirichlet process mixtures: Gibbs sampler (2/2)

- Suppose our target distribution is $p(\mathcal{C}|x_{1:n}) \propto p(x_{1:n}|\mathcal{C})p(\mathcal{C})$.

- Write $\mathcal{C} \setminus i$ for the current partition excluding i.

Gibbs sampler for DPM with conjugate prior

- Start with all customers at the same table: Initialize $\mathcal{C} = \big\{\{1, \ldots, n\}\big\}$.

- For $i = 1, \ldots, n$: Reseat customer $i \ldots$

  at table $c \in \mathcal{C} \setminus i$ with probability $\propto |c| \, \dfrac{m(x_{c \cup i})}{m(x_c)}$,

  at a new table with probability $\propto \alpha \, m(x_i)$

Outline

Applications of DPs and DPMs (1/3)

- Nonparametric model for nuisance distributions in regression, such as:
  - the residual distribution (Kottas & Gelfand, 2001)
  - the distribution of random effects (Bush & MacEachern, 1996; Mukhopadhyay & Gelfand, 1997)
  - errors-in-variables distributions (Müller & Roeder, 1997)

Applications of DPs and DPMs (2/3)

- Building flexible structured models for
  - spatial processes (Gelfand et al., 2005),
  - time-evolving data (Dunson, 2006),
  - conditional density estimation (Dunson et al., 2007),
  - density estimation (Escobar & West, 1995).

Applications of DPs and DPMs (3/3)

- Commonly used for clustering with an unknown number of clusters.
  - e.g., Medvedovic & Sivaganesan (2002), Huelsenbeck & Andolfatto (2007), and many others.

- Flexible model for the component distributions in a mixture model.
  - Rodriguez and Walker (2014)

References and supplements

- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 209-230.
- J. Sethuraman (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4:639–650.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2), 249-265.