Bayesian Nonparametric Density Estimation

version: 2017-11-16 · 19:32:10

Basic Idea

- **Parametric Density Estimation:**

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y \mid \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

  Bayesian inference is on $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{Y}) \times \pi(\boldsymbol{\theta})$.

- **Nonparametric Density Estimation:**

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y), \qquad F \sim \pi(F).$$

  - $\pi(F)$ is a distribution on CDFs.
  - Bayesian inference is on $p(F \mid \mathbf{Y}) \propto \mathcal{L}(F \mid \mathbf{Y}) \times \pi(F)$.
  - **Convenient Prior:** A conjugate prior for $L(F \mid \mathbf{Y})$ is the Dirichlet Process, $F \sim \mathrm{DP}(F_0, \alpha)$.

The Dirichlet Process

Dirichlet Distribution:

- Let $X_k \overset{\text{ind}}{\sim} \text{Gamma}(\alpha\rho_k, \beta)$ for $k = 1, \dots, K$, where $\alpha, \rho_k > 0$ and $\sum_{k=1}^K \rho_k = 1$. Then

$$\mathbf{Y} = \left(\frac{X_1}{\sum_{i=1}^K X_i}, \dots, \frac{X_K}{\sum_{i=1}^K X_i} \right) \sim \text{Dirichlet}(\boldsymbol{\rho}, \alpha).$$

  - $\mathbf{Y}$ is a probability vector: $Y_k > 0$ and $\sum_{k=1}^K Y_k = 1$.
  - Mean and variance:

$$E[\mathbf{Y}] = \boldsymbol{\rho}, \qquad \text{var}(\mathbf{Y}) = \frac{\text{diag}(\boldsymbol{\rho}) - \boldsymbol{\rho}\boldsymbol{\rho}'}{\alpha + 1}.$$

The Dirichlet Process

Dirichlet Process:

- **Notation:** Let $F(y)$ be an arbitrary CDF and $B \subseteq \mathbb{R}$. Then $F(B) := \Pr(Y \in B)$, where $Y \sim F(y)$.
- **Definition:** Let F_0 be a CDF and $\alpha > 0$. Then $F \sim \mathrm{DP}(F_0, \alpha)$ is said to follow a Dirichlet Process if for any finite partition $B_1 \amalg B_2 \amalg \cdots \amalg B_K = \mathbb{R}$,

$$\big(F(B_1), \ldots, F(B_K)\big) \sim \mathrm{Dirichlet}(\boldsymbol{\rho}, \alpha), \qquad \boldsymbol{\rho} = \big(F_0(B_1), \ldots, F_0(B_K)\big).$$

The Dirichlet Process

- **DP:** $F \sim \text{DP}(F_0, \alpha) \iff F(\mathbf{B}) \sim \text{Dirichlet}(F_0(\mathbf{B}), \alpha), \; \forall \; \mathbf{B} = \amalg_{k=1}^{K} B_k = \mathbb{R}$.
- **Representation:** Turns out that $F \sim \text{DP}(F_0, \alpha)$ is a discrete distribution with countably many atoms:

$$F(y) = \textstyle\sum_{k=1}^{\infty} w_k \delta_{y_k}(y).$$

- **Sampling:** Draw $F \sim \text{DP}(F_0, \alpha)$ by stick-breaking procedure:

  1. Draw $y_1, y_2, \ldots \overset{\text{iid}}{\sim} F_0(y)$.

  2. Draw $\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ and let $w_k = \beta_k \prod_{i=1}^{k-1}(1 - \beta_i)$.

  3. $F(y) = \sum_{k=1}^{\infty} w_k \delta_{y_k}(y)$ is a draw from $\text{DP}(F_0, \alpha)$.

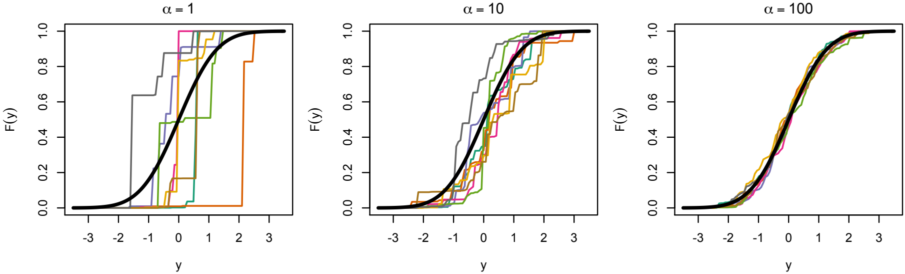
The Dirichlet Process

- **DP:** $F(y) = \sum_{k=1}^{\infty} w_k \delta_{y_k}(y) \sim \mathrm{DP}(F_0, \alpha) \iff F(\mathbf{B}) \sim \mathrm{Dirichlet}(F_0(\mathbf{B}), \alpha)$.
- **Sampling:** Draw $F \sim \mathrm{DP}(F_0, \alpha)$ by stick-breaking procedure:
  1. Draw $y_1, y_2, \ldots \overset{\text{iid}}{\sim} F_0(y)$.
  2. Draw $\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$ and let $w_k = \beta_k \prod_{i=1}^{k-1}(1 - \beta_i)$.
  3. $F(y) = \sum_{k=1}^{\infty} w_k \delta_{y_k}(y)$ is a draw from $\mathrm{DP}(F_0, \alpha)$.
- **In practice:**
  - Can't do this exactly because we can't store an infinite sequence in memory.
  - Instead, draw $F(y) = \sum_{k=1}^{K} w_k \delta_{y_k}(y)$, where K is predetermined by memory allocation, or e.g., note that $w_1 > w_2 > \cdots$, and

$$E[w_k] = \alpha^{k-1}/(1+\alpha)^k,$$

and use this to bound expectation as a function of K. (or use a `while`-loop and stop when $1 - \sum_{k=1}^{K} w_k < \varepsilon$, if dynamic memory allocation is not a concern.)

Example



$F_1, \ldots, F_8 \overset{\text{iid}}{\sim} \text{DP}\{\mathcal{N}(0,1), \alpha\}$ for different values of α.

The Dirichlet Process

- **DP:** $F(y) = \sum_{k=1}^{\infty} w_k \delta_{y_k}(y) \sim \mathrm{DP}(F_0, \alpha) \iff F(\mathbf{B}) \sim \mathrm{Dirichlet}(F_0(\mathbf{B}), \alpha)$.
- **CDF Sampling:** Draw $F(y) \approx \sum_{k=1}^{K} w_k \delta_{y_k}(y)$ by stick-breaking procedure.
- **Marginal Sampling:** Can exactly sample $Y_1, \ldots, Y_n$ from

$$
\begin{aligned}
& Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y), \qquad F \sim \mathrm{DP}(F_0, \alpha) \\
\iff \; & Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \int F(y) \times \pi(F \mid F_0, \alpha) \, \mathrm{d}\{F\}
\end{aligned}
$$

  by Chinese Restaurant Process:

  1. Draw $Y_1 \sim F_0(y)$.
  2. Let $\hat{F}_i(y)$ denote the empirical distribution of $Y_1, \ldots, Y_i$. Draw

$$
Y_{i+1} \sim \frac{\alpha}{\alpha + i} F_0(y) + \frac{i}{\alpha + i} \hat{F}_i(y).
$$

  The Y_i drawn this way are not iid. However, they are exchangeable, i.e., order in which we draw doesn't matter.

The Dirichlet Process

- **DP:** $F(y) = \sum_{k=1}^{\infty} w_k \delta_{y_k}(y) \sim \mathrm{DP}(F_0, \alpha) \iff F(\mathbf{B}) \sim \mathrm{Dirichlet}(F_0(\mathbf{B}), \alpha)$.
- **CDF Sampling:** Draw $F(y) \approx \sum_{k=1}^{K} w_k \delta_{y_k}(y)$ by stick-breaking procedure.
- **Marginal Sampling:**

  Can exactly sample $\quad Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y), \qquad F \sim \mathrm{DP}(F_0, \alpha)$

  by Chinese Restaurant Process:

  1. Draw $Y_1 \sim F_0(y)$.

  2. Let $\hat{F}_i(y)$ denote the empirical distribution of $Y_1, \ldots, Y_i$. Draw

  $$Y_{i+1} \sim \frac{\alpha}{\alpha + i} F_0(y) + \frac{i}{\alpha + i} \hat{F}_i(y).$$

  CRP analogy: Customer $i+1$ enters restaurant, sits at new table with probability $\alpha/(\alpha + i)$ and orders dish $Y_{i+1} \sim F_0(y)$. Otherwise, randomly chooses among existing tables proportionally to how many people are sitting there, and eats whatever dish is at the table.

The Dirichlet Process

- **DP:** $F(y) = \sum_{k=1}^\infty w_k \delta_{y_k}(y) \sim \mathrm{DP}(F_0, \alpha) \iff F(\mathbf{B}) \sim \mathrm{Dirichlet}(F_0(\mathbf{B}), \alpha)$.
- **Marginal Sampling:** Draw $\quad Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F(y), \quad F \sim \mathrm{DP}(F_0, \alpha)$

  with CRP:

  - $Y_{i+1} \sim \frac{\alpha}{\alpha+i} F_0(y) + \frac{i}{\alpha+i} \hat{F}_i(y)$, where $\hat{F}_i(y)$ is the ECDF of $Y_1, \ldots, Y_i$.
- **Marginal Distribution:**
  - Assume that the PDF $f_0(y)$ exists.
  - Let $\tilde{Y}_1, \ldots, \tilde{Y}_K$ denote the unique values of $\mathbf{Y} = (Y_1, \ldots, Y_n)$, and $\mathbf{n} = (n_1, \ldots, n_K)$ denote number of Y_i having each value.
  - The marginal distribution is $\quad p(\mathbf{Y} \mid F_0, \alpha) = \dfrac{\alpha^K \prod_{k=1}^K f_0(\tilde{Y}_k)}{\prod_{i=1}^n (\alpha + i)} \times g(\mathbf{n})$,

    where $g(\mathbf{n})$ doesn't depend on F_0 or α. (combinatorics result on which observations have ties, but interestingly doesn't depend on exact order of ties.)

Bayesian Inference

- **Model and Prior:**

$$\begin{aligned} Y_1, \ldots, Y_n &\overset{\text{iid}}{\sim} F(y) \\ F(y) &\sim \text{DP}\{F_0(y \mid \boldsymbol{\theta}), \alpha\} \\ (\boldsymbol{\theta}, \alpha) &\sim \pi(\boldsymbol{\theta}, \alpha). \end{aligned}$$

- **Conditional Distribution:**

$$F(y) \mid \boldsymbol{\theta}, \alpha, \mathbf{Y} \sim \text{DP}\left\{ \frac{\alpha}{\alpha + n} F_0(y \mid \boldsymbol{\theta}) + \frac{n}{\alpha + n} \underbrace{\hat{F}_n(y)}_{\text{ECDF of } \mathbf{Y}}, \quad \alpha + n \right\}.$$

- **Marginal Likelihood:** Assume PDF $f_0(y \mid \boldsymbol{\theta})$ exists and let $\tilde{Y}_1, \ldots, \tilde{Y}_K$ denote the unique values of $\mathbf{Y}$. Then

$$\mathcal{L}(\boldsymbol{\theta}, \alpha \mid \mathbf{Y}) \propto \frac{\alpha^K \Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{k=1}^{K} f_0(\tilde{Y}_k \mid \boldsymbol{\theta}).$$

(Note that $\prod_{i=1}^{n} (\alpha + i) = \Gamma(\alpha + n)/\Gamma(\alpha)$.)

Dirichlet Process Mixture Model:

- **DPM Model:**

$$
\begin{aligned}
Y_i \mid \boldsymbol{\theta}_i &\overset{\text{ind}}{\sim} f(y \mid \boldsymbol{\theta}_i) \\
\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n &\overset{\text{iid}}{\sim} G(\boldsymbol{\theta}) \\
G(\boldsymbol{\theta}) &\sim \mathrm{DP}\{G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}), \alpha\}
\end{aligned}
$$

- By writing $G(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} w_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$, can view

$$
Y \mid G \sim \sum_{k=1}^{\infty} w_k f(y \mid \boldsymbol{\theta}_k)
$$

as an infinite-component mixture model.

Dirichlet Process Mixture Model

- **DPM Model:**

$$
\begin{aligned}
Y_i \mid \boldsymbol{\theta}_i &\overset{\text{ind}}{\sim} f(y \mid \boldsymbol{\theta}_i) \\
\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n &\overset{\text{iid}}{\sim} G(\boldsymbol{\theta}) \\
G(\boldsymbol{\theta}) &\sim \mathrm{DP}\{G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}), \alpha\}
\end{aligned}
$$

- **Clustering:**
  - $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ takes on $1 \le K \le n$ distinct values $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_K$
  - *Cluster allocation:*

$$
\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{C}}(\mathbf{Y}) = \amalg_{k=1}^{K} \boldsymbol{\mathcal{S}}_k, \qquad \boldsymbol{\mathcal{S}}_k = \{Y_i : \boldsymbol{\theta}_i = \tilde{\boldsymbol{\theta}}_k\}.
$$

  $\Longrightarrow$ don't need to prespecify number of clusters.

  - *Cluster Probability:*

$$
\Pr(\text{cluster allocation is } \boldsymbol{\mathcal{C}}_0 \mid \mathbf{Y}, \boldsymbol{\eta}, \alpha) = \Pr(\boldsymbol{\mathcal{C}}(\mathbf{Y}) = \boldsymbol{\mathcal{C}}_0 \mid \mathbf{Y}, \boldsymbol{\eta}, \alpha)
$$

Dirichlet Process Mixture Model

- **DPM Model:**

$$
\begin{aligned}
Y_i \mid \boldsymbol{\theta}_i &\overset{\text{ind}}{\sim} f(y \mid \boldsymbol{\theta}_i) \\
\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n &\overset{\text{iid}}{\sim} G(\boldsymbol{\theta}) \\
G(\boldsymbol{\theta}) &\sim \text{DP}\{G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}), \alpha\}
\end{aligned}
$$

- **Mixing Kernel:** The most common choice is $Y \mid \mu, \sigma \sim \mathcal{N}(\mu, \sigma^2)$. But to simplify calculations, let

$$
Y \sim \text{NEF}(\boldsymbol{\theta}) \quad \iff \quad \begin{aligned} & f(y \mid \boldsymbol{\theta}) = \exp\{\mathbf{T}'\boldsymbol{\theta} - \Phi(\boldsymbol{\theta}) + h(y)\} \\ & \mathbf{T} = \mathbf{T}(y) \end{aligned}
$$

be a Natural Exponential Family. (Normal is an NEF but using non-standard parametrization.)

Bayesian Inference

- **Model and Prior:**

$$\begin{aligned} Y_i \mid \boldsymbol{\theta}_i &\overset{\text{ind}}{\sim} f(y \mid \boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n &\overset{\text{iid}}{\sim} G(\boldsymbol{\theta}) \\ G(\boldsymbol{\theta}) &\sim \mathrm{DP}\{G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}), \alpha\} \\ (\boldsymbol{\eta}, \alpha) &\sim \pi(\boldsymbol{\eta}, \alpha). \end{aligned}$$

- **Mixing Kernel:** $Y \sim \mathrm{NEF}(\boldsymbol{\theta})$. Usually $Y \sim \mathcal{N}(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- **MCMC Sampling:**
  - $p(\boldsymbol{\eta}, \alpha \mid \boldsymbol{\Theta}, \mathbf{Y}) = p(\boldsymbol{\eta}, \alpha \mid \boldsymbol{\Theta})$, where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, and this is just regular inference for DP.
  - So only need to draw $p(\boldsymbol{\Theta} \mid \boldsymbol{\eta}, \alpha, \mathbf{Y})$ to implement a Gibbs sampler.

Bayesian Inference

- **DPM Model and Prior:**

$$
\begin{aligned}
Y_i \mid \boldsymbol{\theta}_i &\overset{\text{ind}}{\sim} f(y \mid \boldsymbol{\theta}_i) & \qquad G(\boldsymbol{\theta}) &\sim \mathrm{DP}\{G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}), \alpha\} \\
\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n &\overset{\text{iid}}{\sim} G(\boldsymbol{\theta}) & \qquad (\boldsymbol{\eta}, \alpha) &\sim \pi(\boldsymbol{\eta}, \alpha).
\end{aligned}
$$

- **Prior distribution:** Componentwise we have

$$
\boldsymbol{\theta}_i \mid \boldsymbol{\Theta}_{-i}, \boldsymbol{\eta}, \alpha \sim \frac{\alpha}{\alpha + n - 1} G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}).
$$

(Recall that $\boldsymbol{\theta}_i$ from CRP are exchangeable, i.e., order of sampling doesn't matter.)

Bayesian Inference

- **DPM Model and Prior:**

$$Y_i \mid \boldsymbol{\theta}_i \overset{\text{ind}}{\sim} f(y \mid \boldsymbol{\theta}_i) \qquad \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \overset{\text{iid}}{\sim} G(\boldsymbol{\theta}) \qquad G(\boldsymbol{\theta}) \sim \mathrm{DP}\{G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}), \alpha\} \qquad (\boldsymbol{\eta}, \alpha) \sim \pi(\boldsymbol{\eta}, \alpha).$$

- **Prior distribution:**

$$\boldsymbol{\theta}_i \mid \boldsymbol{\Theta}_{-i}, \boldsymbol{\eta}, \alpha \sim \frac{\alpha}{\alpha + n - 1} G_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}).$$

- **Posterior distribution:**

$$\boldsymbol{\theta}_i \mid \boldsymbol{\Theta}_{-i}, \boldsymbol{\eta}, \alpha, \mathbf{Y} \sim r \cdot p(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}, \alpha, Y_i) + \sum_{j \neq i} q_j \cdot \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}), \qquad \text{where}$$

$$q_j \propto \frac{f(Y_i \mid \boldsymbol{\theta}_j)}{\alpha + n - 1}$$

(for each $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_{-i}$: posterior $\propto$ prior $\times$ likelihood)

$$r \propto \int f(Y_i \mid \boldsymbol{\theta}) \frac{\alpha \cdot g_0(\boldsymbol{\theta} \mid \boldsymbol{\eta})}{\alpha + n - 1} \, \mathrm{d}\boldsymbol{\theta}$$

$$\left(\begin{aligned} r &= \Pr(\boldsymbol{\theta}_i \notin \boldsymbol{\Theta}_{-i} \mid \boldsymbol{\eta}, \alpha, \mathbf{Y}) \\ &\propto p(Y_i \mid \boldsymbol{\eta}) \times \alpha/(\alpha + n - 1) \\ &= \tfrac{\alpha}{\alpha+n-1} \textstyle\int f(Y_i \mid \boldsymbol{\theta}) g_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\theta} \end{aligned} \right)$$

$$r + \sum_{j \neq i} q_j = 1.$$

Bayesian Inference

- **DPM Posterior distribution:**

$$\boldsymbol{\theta}_i \mid \boldsymbol{\Theta}_{-i}, \boldsymbol{\eta}, \alpha, \mathbf{Y} \sim r \cdot p(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}, Y_i) + \sum_{j \neq i} q_j \cdot \delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}),$$

$q_j \propto f(Y_i \mid \boldsymbol{\theta}_j)$
$r \propto \alpha \int f(Y_i \mid \boldsymbol{\theta}) g_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\theta}$
$r + \sum_{j \neq i} q_j = 1$

- **Calculation:** Difficult in most cases. But when

$$Y \mid \boldsymbol{\theta} \sim \mathrm{NEF}(\boldsymbol{\theta}) \quad \iff \quad f(y \mid \boldsymbol{\theta}) = \exp\{\mathbf{T}'\boldsymbol{\theta} - \Phi(\boldsymbol{\theta}) + h(y)\}$$

and g_0 is the conjugate prior

$$\boldsymbol{\theta} \mid \boldsymbol{\eta} \sim \mathrm{cEF}(\boldsymbol{\Upsilon}, \nu) \quad \iff \quad g_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = \exp\{\boldsymbol{\Upsilon}'\boldsymbol{\theta} - \nu\Phi(\boldsymbol{\theta}) + q(\boldsymbol{\Upsilon}, \nu)\},$$

then $p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, Y) = g_0(\boldsymbol{\theta} \mid \boldsymbol{\Upsilon} + \mathbf{T}, \nu + 1)$, and

$$\int f(Y \mid \boldsymbol{\theta}) g_0(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \, \mathrm{d}\boldsymbol{\theta} = p(Y \mid \boldsymbol{\eta}) = \frac{f(Y \mid \boldsymbol{\theta}) g_0(\boldsymbol{\theta} \mid \boldsymbol{\eta})}{p(\boldsymbol{\theta} \mid \boldsymbol{\eta}, Y)} = \frac{\exp\{h(Y) + q(\boldsymbol{\Upsilon}, \nu)\}}{\exp\{q(\boldsymbol{\Upsilon} + \mathbf{T}, \nu + 1)\}}.$$

$\implies$ r can also be calculated explicitly.