BNPlib for density estimation

A new nonparametric C++ library

Bruno Guindani
Elena Zazzetti

POLITECNICO
MILANO 1863

November 22, 2019

https://github.com/poliprojects/BNPlib

Non-Parametric statistics

- Goal: density estimation
- **Infinite-dimensional** parameters, e.g. functions

Non-Parametric statistics (the Bayesian way)

- Goal: density estimation
- **Infinite-dimensional** parameters, e.g. functions
- Model:

$$
\begin{aligned}
y_i | G &\overset{\text{iid}}{\sim} G, \quad i = 1, \dots, n \\
G &\sim \mathcal{P}
\end{aligned}
$$

Non-Parametric statistics (the Bayesian way)

- Goal: density estimation
- **Infinite-dimensional** parameters, e.g. functions
- Model:

$$
\begin{aligned}
y_i | G &\overset{\text{iid}}{\sim} G, \quad i = 1, \dots, n \\
G &\sim \mathcal{P}
\end{aligned}
$$

$$
\mathcal{P} : \Omega \to M(S) \;\; \text{fixed}
$$

$$
\left[\; \omega \mapsto G(\cdot) \;\right]
$$

- Model name: **BNP model**

Dirichlet Process prior

$$\begin{aligned} y_i | G &\overset{\text{iid}}{\sim} G \\ G &\sim \mathcal{P} = DP(MG_0) \end{aligned}$$

- Parameters: $M > 0,\ G_0 \in M(S)$
- Defining property: $\forall \{B_{1:k}\}$ partition of S,

$$[G(B_1), \ldots, G(B_k)] \sim \mathrm{Dir}\left(MG_0(B_1), \ldots, MG_0(B_k)\right)$$

Dirichlet Process prior

$$y_i|G \overset{\text{iid}}{\sim} G$$
$$G \sim \mathcal{P} = DP(MG_0)$$

- Parameters: $M > 0,\ G_0 \in M(S)$
- Defining property: $\forall\{B_{1:k}\}$ partition of S,

$$[G(B_1), \dots, G(B_k)] \sim \text{Dir}\left(MG_0(B_1), \dots, MG_0(B_k)\right)$$

- **Discreteness** (stick-breaking): $G(\cdot) = \sum_{k=1}^{+\infty} w_h \delta_{m_h}(\cdot)$
- **Conjugacy**: $G|\mathbf{y} \sim DP(MG_0 + \sum_i \delta_{y_i}) \quad \Longrightarrow$ density estimation

Continuous density estimation

- **Mixtures** (kernel f + mixing distribution G):

$$y_i | G \sim f_G(y) = \int f_\vartheta(y) \, \mathrm{d}G(\vartheta)$$

Continuous density estimation

- **Mixtures** (kernel f + mixing distribution G):

$$y_i|G \sim f_G(y) = \int f_\vartheta(y)\,\mathrm{d}G(\vartheta)$$
$$G \sim DP(MG_0)$$

- Model name: **DPM model**

Continuous density estimation

- **Mixtures** (kernel f + mixing distribution G):

$$y_i|G \sim f_G(y) = \int f_\vartheta(y)\,\mathrm{d}G(\vartheta)$$
$$G \sim DP(MG_0)$$

- Model name: **DPM model**
- Equivalent to:

$$y_i|\vartheta_i \overset{\perp\!\!\!\perp}{\sim} f_{\vartheta_i}$$
$$\vartheta_i|G \overset{\text{iid}}{\sim} G$$
$$G \sim DP(MG_0)$$

- ϑ_i "latent variables" $\forall i = 1, \ldots, n$

Clustering in the DPM

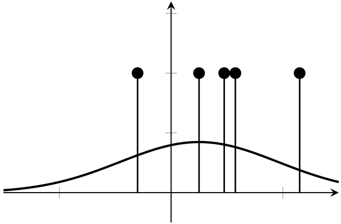
- Discreteness: the ϑ_i have one of the k **unique values** ϕ_j $_{(j\,=\,1,\ldots,k)}$
- $k \simeq M \log(n) \ll n$
- All i s.t. $\vartheta_i = \phi_j$ belong to cluster S_j $_{(j\,=\,1,\ldots,k)}$, and $n_j = |S_j|$

Clustering in the DPM

- Discreteness: the ϑ_i have one of the k **unique values** ϕ_j $_{(j\,=\,1,\ldots,k)}$
- $k \simeq M\log(n) \ll n$
- All i s.t. $\vartheta_i = \phi_j$ belong to cluster S_j $_{(j\,=\,1,\ldots,k)}$, and $n_j = |S_j|$
- Conditional prior for ϑ_i:

$$\mathcal{L}(\vartheta_i|\boldsymbol{\vartheta}_{-i}) \propto \sum_{j=1}^{k^-} n_j^- \delta_{\phi_j^-}(\vartheta_i) + MG_0(\vartheta_i)$$

$$\uparrow \qquad\qquad\qquad \uparrow$$

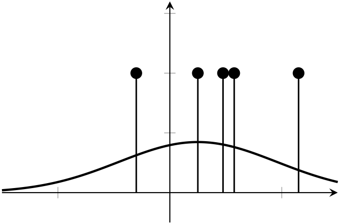


Clustering in the DPM

- Discreteness: the ϑ_i have one of the k **unique values** ϕ_j $_{(j\,=\,1,\ldots,k)}$
- $k \simeq M \log(n) \ll n$
- All i s.t. $\vartheta_i = \phi_j$ belong to cluster S_j $_{(j\,=\,1,\ldots,k)}$, and $n_j = |S_j|$
- Conditional prior for ϑ_i:

$$\mathcal{L}(\vartheta_i|\boldsymbol{\vartheta}_{-i}) \propto \sum_{j=1}^{k^-} n_j^- \delta_{\phi_j^-}(\vartheta_i) + M G_0(\vartheta_i)$$

$$\uparrow \qquad\qquad\qquad \uparrow$$



- Conditional posterior for ϑ_i:

$$\mathcal{L}(\vartheta_i|\boldsymbol{\vartheta}_{-i}, y_i) \propto \sum_{j=1}^{k^-} f_\vartheta(y_i)\delta_{\phi_j^-}(\vartheta_i) + M\, r_i\, G_0(\vartheta_i|y_i)$$

Neal's Algorithm 2

Equivalent models as $K \to +\infty$:

$$
\begin{aligned}
(Y_i|\vartheta_i) &\sim F(\vartheta_i)\\
(\vartheta_i|G) &\sim G\\
G &\sim DP(M, G_0)
\end{aligned}
$$

(hierarchical model)

$$
\begin{aligned}
(Y_i|\phi, c_i) &\sim F(\phi_{c_i})\\
(c_i|\mathbf{p}) &\sim \text{Discrete}(p_1, ..., p_K)\\
\phi_c &\sim G_0\\
\mathbf{p} &\sim \text{Dirichlet}(M/K, ..., M/K)
\end{aligned}
$$

(discrete model)

with $c_i = j \iff \vartheta_i \in S_j$ allocation parameters and

$$\boldsymbol{\vartheta} \leftrightsquigarrow (\boldsymbol{\phi}, \mathbf{c})$$

Neal's Algorithm 2

Algorithm:

- For $i = 1...n$: update c_i
  - If c_i is associated with no other observation, remove ϕ_{c_i} from the state
  - Gibbs sampling for c_i:

$$\text{If } c = c_j \text{ for some } j \neq i: \; P(c_i = c | c_{-i}, y_i, \phi) \propto \frac{n_{-i,c}}{n-1-M} F(y_i, \phi_c)$$

$$P(c_i \neq c_j \text{ for all } j | c_{-i}, y_i, \phi) \propto \frac{M}{n-1-M} \int F(y_i, \phi) dG_0(\phi)$$

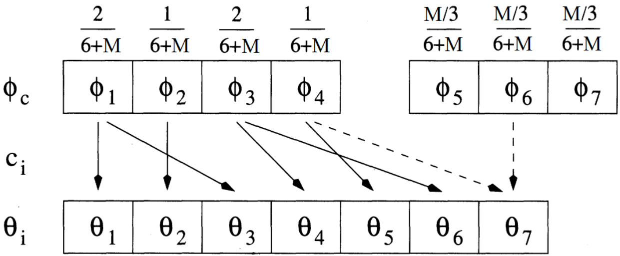
  - If the new c_i is not associated with any other observation, draw a value for ϕ_{c_i} from H_i and add it to the state
- For $c \in \{c_1, .., c_n\}$: update ϕ_c given y_i such that $c_i = c$

Advantages

- Feasible if we can compute $\int F(y_i, \phi) dG_0(\phi)$ and sample from H_i (generally conjugate case)
- Change the ϑ for more than one observation simultaneously

Neal's Algorithm 8

- Approach:
  - Markov chain with permanent state **c** and ϕ
  - Gibbs sampling to the state extended by the addition of m auxiliary parameters



- Prior for c_i:

If $c = c_j$ for some j : $P(c_i = c | c_{-i}) = \dfrac{n_{-i,c}}{n-1-M}$

$$P(c_i \neq c_j \text{ for all } j) = \frac{M}{n-1-M} \Rightarrow$$ **split** among the m auxiliary parameters

Neal's Algorithm 8

Algorithm:

- For $i = 1...n$: update c_i
  - Sample auxiliary parameters:
    - $c_i = c_j$ for some j $\Rightarrow$ no connection
    - $c_i \neq c_j$ $\Rightarrow$ association to one of m

    The other ϕ values drawn from G_0
  - Gibbs sampling update for c_i:

$$P(c_i = c|c_{-i}, y_i, \phi_1, ..., \phi_h) \propto \begin{cases} \frac{n_{-i,c}}{n-1-M} F(y_i, \phi_c), & \text{for } 1 \leq c \leq k^- \\ \frac{M/m}{n-1-M} F(y_i, \phi_c), & \text{for } k^- + 1 < c \leq h \end{cases}$$

  - Discard ϕ values not associated
- For $c \in \{c_1, .., c_n\}$: update ϕ_c given y_i such that $c_i = c$

Advantages

- Models with non-conjugate priors
- As $m \to +\infty$ it approaches Algorithm 2 but equilibrium distribution is exact
- More efficient than similar algorithms (e.g. no-gaps)
- Hierarchical extensions

Stick-Breaking Priors

$$\mathscr{P}(\cdot) = \sum_{k=1}^{N} p_k \delta_{Z_k}(\cdot)$$

$$p_k = V_1 \text{ and } p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1})V_k$$

$$\mathbf{Z_k} \overset{\text{iid}}{\sim} H$$

$$V_k \overset{\text{iid}}{\sim} Beta(a_k, b_k)$$

$$\mathbf{a} = (a_1, a_2, ...) \text{ and } \mathbf{b} = (b_1, b_2, ...)$$

$$0 \leq p_k \leq 1 \text{ and } \sum_{k=1}^{N} p_k = 1$$

- $N < +\infty$: $\mathscr{P}_N(\mathbf{a}, \mathbf{b})$
  - $\mathbf{p} \sim \mathscr{GD}(\mathbf{a}, \mathbf{b})$
  - e.g. all finite dimensional Dirichlet priors
- $N = +\infty$: $\mathscr{P}_\infty(\mathbf{a}, \mathbf{b})$
  - e.g. Dirichlet process, the two parameter Poisson-Dirichlet process

Blocked Gibbs Algorithm

- Assumption: **finite-dimensional** prior $P \sim \mathscr{P}_N(\mathbf{a}, \mathbf{b})$
- Finite number of variables $\Rightarrow$ *blocks of parameters*
- **Model**:

$$
\begin{aligned}
(Y_i | \phi, \mathbf{c}) &\overset{\perp\!\!\!\perp}{\sim} F(\phi_{c_i}),\ i = 1, .., n \\
(c_i | \mathbf{p}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k \delta_k(\cdot) \\
\mathbf{p} &\sim \mathscr{G}\mathscr{D}(\mathbf{a}, \mathbf{b}) \\
\phi_c &\sim G_0
\end{aligned}
$$

Blocked Gibbs Algorithm

Algorithm:
Repeatedly drawing values from conditional distributions of the blocked variables:

- $(\phi|\mathbf{c}, \mathbf{Y})$
- $(\mathbf{c}|\phi, \mathbf{p}, \mathbf{Y})$
- $(\mathbf{p}|\mathbf{c})$

Direct sampling of the posterior $\mathscr{P}(\cdot|\mathbf{Y})$:

- The Algorithm produces draws from $(\phi, \mathbf{c}, \mathbf{p}|\mathbf{Y})$
- Each draw $(\phi, \mathbf{c}, \mathbf{p})$ defines a measure $P(\cdot) = \sum_{k=1}^{N} p_k \delta_{\phi_k}(\cdot)$
- Each P is a drawn from $\mathscr{P}(\cdot|\mathbf{Y})$

Advantages

- Handles the issue of conjugacy
- Good mixing
- Hierarchical extensions

Bibliography

- Muller, Quintana, *Bayesian Nonparametric Data Analysis*
- Neal (2000), *Markov Chain Sampling Methods for Dirichlet Process Mixture Models*
- Ishwaran, James (2001), *Gibbs Sampling Methods for Stick-Breaking Priors*