Title

Subtitle

Bruno Guindani
Elena Zazzetti

November 22, 2019

https://github.com/poliprojects/BNPlib

Title

Non-Parametric statistics

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

Non-Parametric statistics (the Bayesian way)

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

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

Non-Parametric statistics (the Bayesian way)

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

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

$$\mathcal{P} : \Omega \to M(S) \ \ \text{fixed}$$
$$\left[\ \omega \mapsto G(\cdot)\ \right]$$

- Model name: **BNP model**

Dirichlet Process prior

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

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

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

Dirichlet Process prior

$$y_i|G \stackrel{\text{iid}}{\sim} G$$
$$G \sim DP(MG_0)$$

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

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

- **Discreteness** (stick-breaking): $G(\cdot) = \sum_{k=1}^{+\infty} w_h \delta_{m_h}(\cdot)$
- Polya urn representation:

$$\mathcal{L}(y_i|y_1, \ldots, y_{i-1}) \propto \sum_{h=1}^{i-1} \delta_{y_h}(y_i) + MG_0(y_i)$$

- $M \log n \ll n$ values

Continuous density estimation

- **Mixtures**:

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

Continuous density estimation

- **Mixtures**:

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

- Model name: **DPM model**

Continuous density estimation

- **Mixtures**:

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

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

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

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

Clustering

- Discreteness: the ϑ_i have one of the k **unique values** ϑ_j^* $(j = 1, \ldots, k)$
- $k \simeq M \log(n) \ll n$ (!)
- All i s.t. $\vartheta_i = \vartheta_j^*$ belong to cluster S_j $\quad (j = 1, \ldots, k)$
-
$$\mathcal{L}(\vartheta_i | \vartheta_i) \propto \sum_{j=1}^{k^-} n_j^- \delta_{\vartheta_j^{*-}}(\vartheta_i) + M G_0(\vartheta_i)$$

Title

Title

Stuff

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*