Conditional Predictive Sampler for Dirichlet process mixture model

Riccardo Corradin - joint with Antonio Canale and Bernardo Nipoti
Dept. of Statistics and Quantitative methods, University of Milano-Bicocca, Italy
Dept. of Computer Science and Statistics, Trinity College Dublin, Ireland
riccardo.corradin@unimib.it

Main idea
Consider a Dirichlet process mixture models (DPM): let

\[ X_i | \theta_i \sim k(X_i; \theta_i) \quad i = 1, \ldots, n \]
\[ \theta_i | \beta \sim \tilde{P} \quad i = 1, \ldots, n \]

with \( k(\cdot, \cdot) \) kernel, \( \alpha \) total mass and \( P_0 \) base measure. We want to estimate \( f(x) \) density of \( X \).

The main idea behind conditional algorithms for DPM models is to sampling only finite-dimensional summaries, possibly of random dimension, of the infinite-dimensional random object \( \beta \). The known conditional methods are based on a stick-breaking representation: truncated SB approach [1], slice sampler approach [2] and retrospective sampler approach [3]. We devise an algorithm based on a conditional version of the predictive distribution and which exploit the peculiar properties of the Dirichlet Process (DP).

Properties of the DP

- **Conjugacy** If \( \theta_i \sim \tilde{P} \) and \( \beta \) is a DP, then we have that \( \tilde{P} | \theta_i \) is still a DP with updated parameters.

- **Finite-dimensional distribution** A DP evaluated on a finite partition of the support follows a Dirichlet distribution.

- **Self-similarity** Given a measurable set \( A \) such that \( 0 < P_0(A) < 1 \), the random probability measure \( \tilde{P} | A \) is independent of \( \tilde{P}(A) \) and \( \tilde{P}(A^c) \). Moreover \( \tilde{P} | A \) is still a DP with total mass \( \alpha P_0(A) \) and base measure \( P_0 | A \).

Methodology
Let \( X = (X_1, \ldots, X_n) \) be the vector of observations, \( \theta = (\theta_1, \ldots, \theta_n) \) vector of parameters, \( \theta^* = (\theta_1^*, \ldots, \theta_n^*) \) their unique values, \( (n_1, \ldots, n_n) \) their frequencies, and \( \Theta^* = \Theta \setminus \{\theta_1^*, \ldots, \theta_n^*\} \). We call conditional predictive distribution (reminds of [4])

\[ P(\theta_i \in dt | X, \beta) \propto \tilde{P}(\theta^*) k(X_i, t) \tilde{q}(dt) + \sum_{j=1}^{k} \tilde{p}(\theta_j^*) k(X_j, t) \tilde{k}_j(dt) \quad i = 1, \ldots, n \]

where \( \tilde{q} = \tilde{P}(\Theta^*) \). By exchangeability, conditionally on \( \tilde{P} \), the \( \theta_i \) are independent: the \( \theta_j^* \) in (1) are atoms in the base measure of \( \tilde{P} \). By conjugacy and finite dimensional distribution we have that \( \tilde{p}_0(X), \tilde{p}_j(X) \), the probabilities vector of \( \Theta^* \), follows a Dirichlet distribution. By self-similarity, \( \tilde{q} \) is a DP. To evaluate \( k(X, \theta^*) \tilde{q}(dt) \), an \( m \)-size i.i.d. sample is taken from \( \tilde{q} \) and weighted by means of the kernel, inspired by [5].

For each iteration \( r = 1, \ldots, n_{\text{iter}} \):

- **Step 1** For each non-empty cluster, update the parameters as
  \[ P(\theta_{jr}^{(r-1)} \in dt | X, \beta) \propto P_0(dt) \prod_{i=1}^{m} \tilde{p}(\theta_j^*) \tilde{k}_j(dt) \]

- **Step 2** Sample \( \theta_{jr}^{(r)} \sim \text{Dir}(\alpha, n_j^{(r-1)}, \ldots, n_k^{(r-1)}) \)

- **Step 3** Sample the m-dimensional vector \( \theta_{\text{temp}}^{(r)} \) from \( \tilde{q} \), by using the Blackwell-MacQueen Polya urn scheme [4].

- **Step 4** For each \( i = 1, \ldots, n \) sample
  \[ P(\theta_i^{(r)} = t | \beta_{jr}^{(r)}, \theta_{\text{temp}}^{(r)}, \theta_j^{(r-1)}, X) = \left\{ \begin{array}{ll}
  \frac{1}{m} \tilde{p}(\theta_i^*) \sum_{k=1}^{m} k(X_i, \theta_k^{(r-1)}) & \text{if } t = \theta_i^{(r-1)}, \ldots, \theta_k^{(r-1)} \\
  \sum_{k=1}^{m} \tilde{p}(\theta_i^*) \tilde{k}(X_i, \theta_k^{(r-1)}) & \text{if } t = \theta_i^{(r-1)}, \ldots, \theta_k^{(r-1)}
\end{array} \right. \]

Simulations results
Toy example we simulated a dataset from a mixture of two Gaussian distribution

\[ X \sim \frac{1}{3} N(-2.5,1) + \frac{2}{3} N(2.5,1) \]

We estimated the model via conditional predictive sampler and via Slice sampler.

Remarks and future
The proposed method combine the intuitive simplicity of the predictive distribution of a DP with a conditional approach. Due to the nature of conditional methods, it is strictly faster than the equivalent marginal methods and it is possible to parallelize the allocation of observations. This make DPM models amenable of use in the big sample size context. Importantly, the method shows desirable properties in terms of performance and quality of the generated samples.

Our purpose for the future are:

- Investigate the possibility of extending the methodology to other processes.
- Extend the model considering hyperprior distributions on the main parameter of the base measure and on the mass of the process.

References