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

$$X_i|\theta_i \overset{ind}{\sim} k(X_i,\theta_i) \qquad i=1,\dots,n$$
$$\theta_i|\tilde{p} \overset{iid}{\sim} \tilde{p} \qquad i=1,\dots,n$$
$$\tilde{p} \sim DP(\alpha P_0)$$

with $k(\cdot,\cdot)$ kernel, α 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 $\tilde{p}$. 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 Procss (**DP**).

Properties of the DP

Conjugacy If $\theta_i \overset{ind}{\sim} \tilde{p}$ and $\tilde{p}$ is a DP, then we have that $\tilde{p}|\theta_1,\dots,\theta_n$ 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 $\mathbf{X} = (X_1,\dots X_n)$ be the vector of observations, $\theta = (\theta_1,\dots,\theta_n)$ vector of parameters, $\theta^* = (\theta_1^*,\dots,\theta_k^*)$ their unique values, $(n_1,\dots,n_k)$ their frequencies, and $\Theta^* = \Theta \setminus (\theta_1^*,\dots,\theta_k^*)$. We call conditional predictive distribution (reminds of [4])

$$P[\theta_i \in dt|\mathbf{X},\tilde{p}] \propto \tilde{p}(\Theta^*)k(X_i,t)\tilde{q}(dt) + \sum_{j=1}^{k} \tilde{p}(\theta_j^*)k(X_i,\theta_j^*)\delta_{\theta_j^*}(dt) \qquad i=1,\dots,n \qquad (1)$$

where $\tilde{q} = \tilde{p}|_{\Theta^*}$. By exchangeability, conditionally on $\tilde{p}$, the θ_i are independent: the θ_j^* in (1) are atoms in the base measure of $\tilde{p}$. By conjugacy and finite dimensional distribution we have that $\tilde{p}_{\Theta^*,\theta^*}$, the probabilities vector of $\{\Theta^*,\theta^*\}$, follows a Dirichlet distribution. By self-similarity, $\tilde{q}$ is a DP. To evaluate $k(X_i,\theta)\tilde{q}(d\theta)$, 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 $r = 1,\dots,n_{iter}$:

- **Step 1** For each non-empty cluster, update the parameters as

$$P\left[\theta_j^{*(r-1)} \in dt|\mathbf{X},\theta\right] \propto P_0(dt)\prod_{i:\theta_i^{(r-1)}=\theta_j^{*(r-1)}} k(X_i;t)$$

- **Step 2** Sample $\tilde{p}^{(r)}_{\Theta^*,\theta^*} \sim \text{Dir}(\alpha, n_1^{(r-1)},\dots,n_k^{(r-1)})$
- **Step 3** Sample the m-dimensional vector $\theta^{(r)}_{\text{temp}}$ from $\tilde{q}$, by using the Blackwell-McQueen Pólya urn scheme [4].
- **Step 4** For each $i = 1,\dots,n$ sample

$$P\left[\theta_i^{(r)} = t \mid \tilde{p}_{\Theta^*,\theta^*},\theta^{(r)}_{\text{temp}},\theta^{*(r-1)},\mathbf{X}\right] = \begin{cases} \frac{1}{m}\tilde{p}^{(r)}(\Theta^*)\sum_{l=1}^{m} k(X_i;\theta^{(r)}_{\text{temp},l}) & \text{if } t = \theta^{(r)}_{\text{temp},1},\dots,\theta^{(r)}_{\text{temp},m} \\ \sum_{j=1}^{k}\tilde{p}^{(r)}(d\theta_j^{*(r-1)})k(X_i;\theta_j^{*(r-1)}) & \text{if } t = \theta_1^{*(r-1)},\dots,\theta_k^{*(r-1)} \end{cases}$$

Simulation 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.

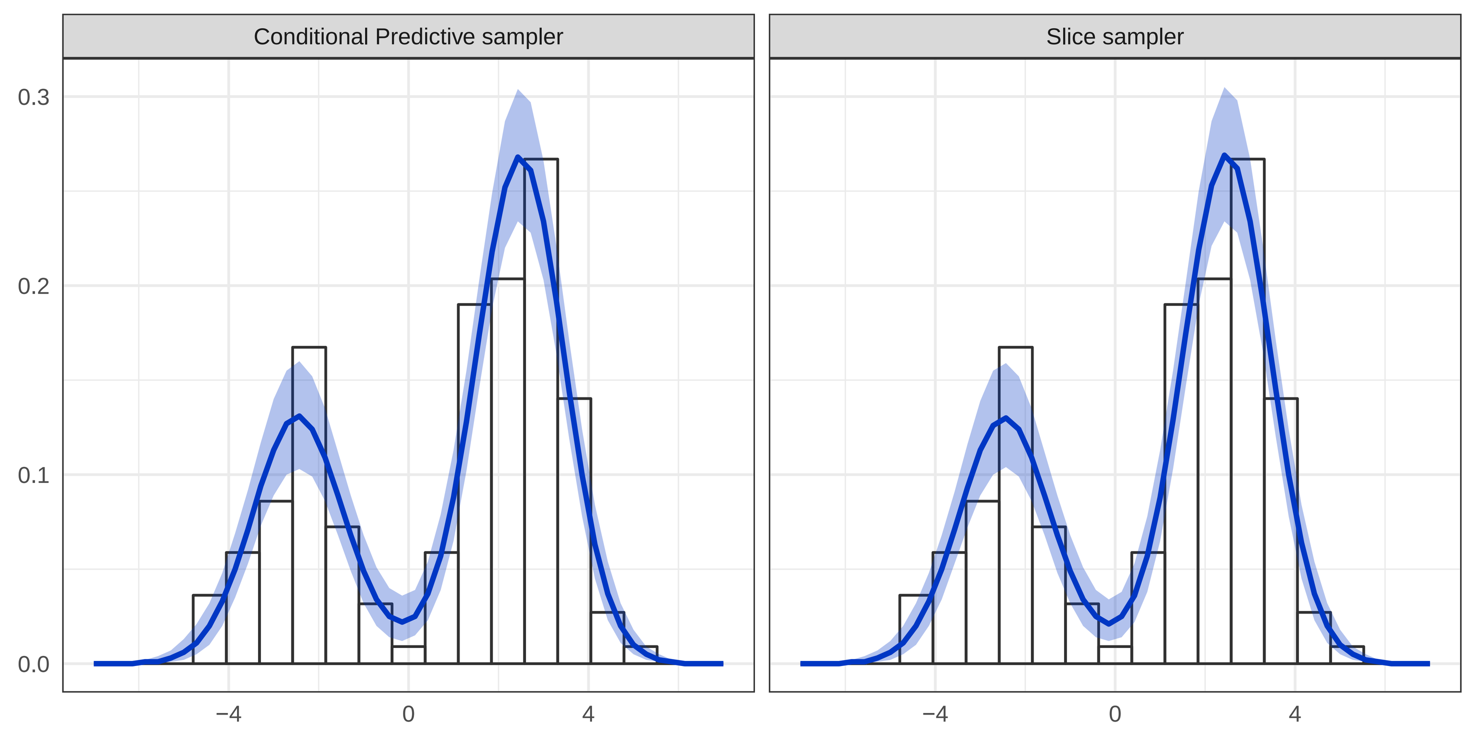

Fig. 1: Estimated density of one replication with 10 · 000 iterations after 2 · 000 burn-in.

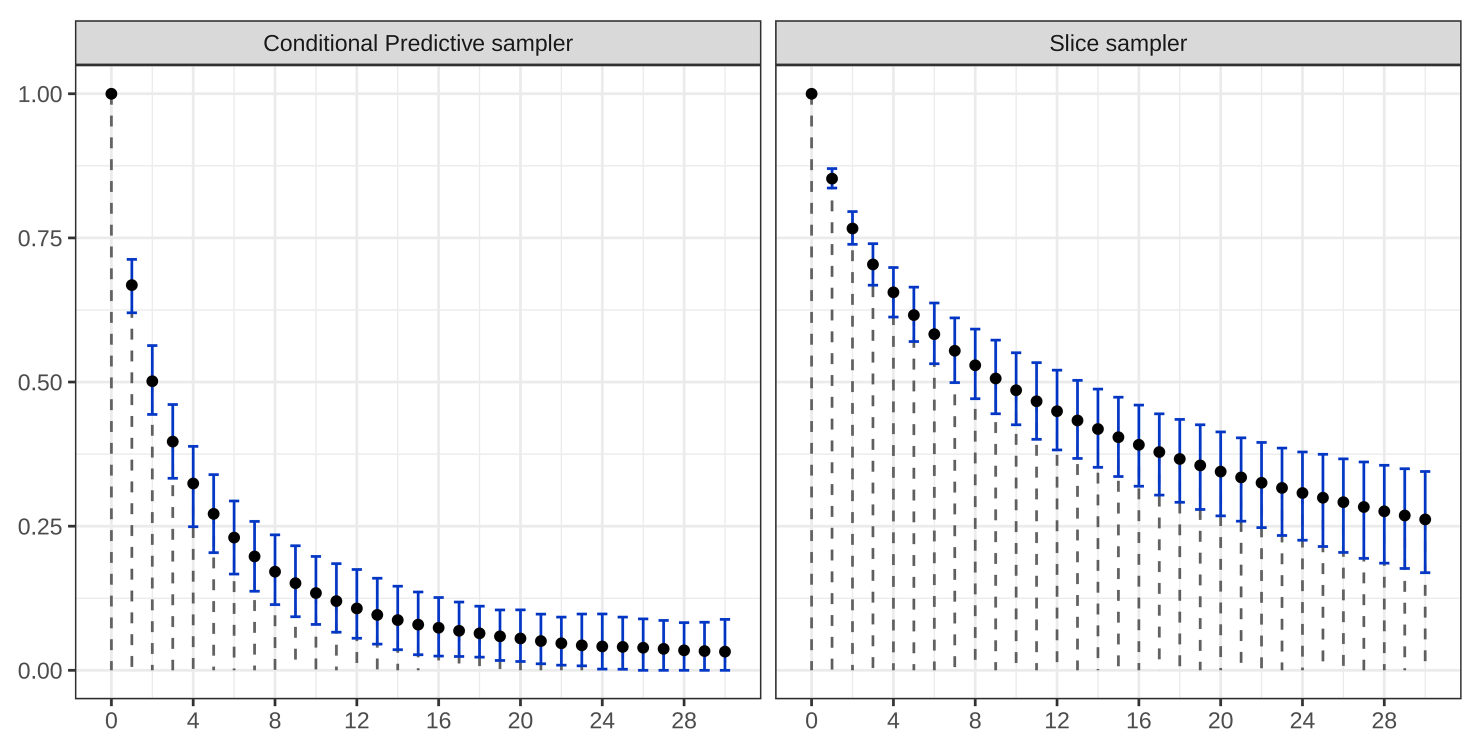

Fig. 2: Estimated ACF between number of groups (averaged over 100 of replications).

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 desirables 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

[1] Ishwaran, H. and Zarepour, M.: *Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models*, Biometrika, 2000.

[2] Walker, S. G.: *Sampling the Dirichlet Mixture Model with Slices*, Communications in Statistics - Simulation and Computation, 2007.

[3] Papaspiliopoulos, O. and Roberts, G. O.: *Retrospective Markov chain Monte Carlo methods for Dirichlet process hirerarchial models*, Biometrika, 2008.

[4] Blackwell, D., MacQueen, J.B.: *Ferguson Distributions Via Polya Urn Schemes*, The Ann. of Statistics, 1973.

[5] Neal, R. M. *Markov Chain Sampling Methods for Dirichlet Process Mixture Models*, Journal of Computational and Graphical Statistics, 2000.