Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. Supplementary material

Sophie Donnet, Judith Rousseau, Vincent Rivoirard, Catia Scricciolo

Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures Supplementary material

Sophie Donnet*, Vincent Rivoirard†, Judith Rousseau‡and Catia Scricciolo§

June 16, 2014

Abstract

1 Gibbs algorithm

We detail the algorithm used to sample from the posterior distribution $(\overline{\lambda}, A, \gamma)|N$ in the Poisson process context, in the most complete case (with a hierarchical level on γ). In case where γ is set to a fixed value, then the corresponding part in the algorithm is removed. As a standard Gibbs algorithm, the simulation is decomposed into three steps:

$$[1] \quad \overline{\lambda}|A, \gamma, N \qquad [2] \quad A|\overline{\lambda}, \gamma, N \qquad [3] \quad \gamma|A, \overline{\lambda}, N.$$

where N is the observed Poisson process over $[0, T]$, namely a number of jumps $N(T)$ and jump instants $(T_1, \ldots, T_{N(T)})$. In order to avoid an artificial truncation in $\overline{\lambda}$, we use the slice sampler strategy proposed by Fall and Barat (2012). More precisely, we consider the stick breaking representation of $\overline{\lambda}$. Let c_i be the affectation variable of data W_i. The DPM model is written as:

$$W_i|c_i, \boldsymbol{\theta}^* \sim h_{\theta^*_{c_i}}, \quad P(c_i = k) = w_k, \forall k \in \mathbb{N}^* \quad (w_k)_{k\in\mathbb{N}^\star} \sim Stick(A), \quad (\theta^*_k)_{k\in\mathbb{N}^\star} \sim_{i.i.d} G_\gamma.$$

The slice sampler strategy consists in introducing a latent variable u_i such that the joint distribution of (W_i, u_i) is $p(W_i, u_i|\omega, \boldsymbol{\theta}^*) = \sum_{k=1}^{\infty} w_k h_{\theta^*_k}(W_i) \frac{1}{\xi_k} \mathbb{1}_{[0,\xi_k]}(u_i)$ with $\xi_k = \min(w_k, \zeta)$, which can be reformulated as:

$$p(W_i, u_i|\omega, \boldsymbol{\theta}^*) = \frac{1}{\zeta}\mathbb{1}_{[0,\zeta]}(u_i) \sum_{k=1, w_k>\zeta}^{\infty} w_k h_{\theta^*_k}(W_i) + \sum_{k=1, u_i\leq w_k\leq\zeta}^{\infty} h_{\theta^*_k}(W_i)\mathbb{1}_{[0,w_k]}(u_i) \quad (1.1)$$

$(w_k)_{k\geq 1}$ verifying $\sum_{k\geq 1} w_k = 1$ (implying $\lim_{k\to\infty} w_k = 0$), the cardinal of $\{k, w_k > \varepsilon\}$ is finite for every $\varepsilon > 0$, and the sum in (1.1) is finite.

Remark 1 *Note that the number of non-null terms in (1.1) is rigorously dependent of the observation index i (denoted K^*_i). But in the algorithm we will deal with the maximum of the K^*_i:*

$$K^* = \max\{K^*_i, i = 1 \ldots N(T)\}$$

The Gibbs algorithm with the Slice sampler strategy now takes into account the latent variable $\mathbf{u} = (u_1, \ldots, u_{N(T)})$ which is sampled conjointly with $\overline{\lambda}$, resulting into the following steps:

$$[1^\star] \quad \overline{\lambda}, \mathbf{u}|A, \gamma, N \qquad [2^\star] \quad A|\overline{\lambda}, \mathbf{u}, \gamma, N \qquad [3^\star] \quad \gamma|A, \overline{\lambda}, \mathbf{u}, N.$$

We now detail steps $[1^\star]$, $[2^\star]$ and $[3^\star]$.

*Université de Paris Dauphine, France
†Université de Paris Dauphine, France
‡CREST, France
§Bocconi University, Italy

1.1 Details of the algorithm

Initialisation The Gibbs algorithms are initialized on $A^{(0)} = 10$ (0 referring to the iteration number of the Gibbs algorithm). We set: $K^{(0)} = N_T$. for $k = 1 \dots K$, $(\theta^*)_k^{(0)} \sim G_\gamma$. When the hierarchical approach is considered on γ ($\gamma \sim \Gamma(a_\gamma, b_\gamma)$), we initialize γ on its prior mean value: $\gamma^{(0)} = \frac{a_\gamma}{b_\gamma}$.

[1*] Sampling from $\overline{\lambda}, \mathbf{u}|A, \gamma, N$

Step [1*] of the Gibbs algorithm is decomposed into 5 steps which are detailed below. Let $\mathbf{u} = (u_1, \dots, u_{N(T)})$, $\mathbf{c} = (c_1, \dots, c_{N(T)})$, $\boldsymbol{\theta}^* = (\theta_1, \dots, \theta_{K^\star})$, $\omega = (w_1, \dots, w_{K^\star})$ –$\mathbf{c}$, $\boldsymbol{\theta}^*$ and ω representing $\overline{\lambda}$– be the current object. We denote by $K_{N(T)}$ the number of non-empty classes:

$$K_{N(T)} = \#\{k \in \{1 \dots K^\star\} | \exists j \in \{1, \dots, N(T)\} \text{ such that } c_j = k\}$$

$\boldsymbol{\theta}^*$ and ω are ordered such that the elements indexed from $K_{N(T)} + 1$ to $K^\star$ correspond to empty classes. $\mathbf{u}$, $\mathbf{c}$, $\boldsymbol{\theta}^*$ and ω are iteratively sampled as follows:

[1*.a] First we sample $\mathbf{u}|\omega, \boldsymbol{\theta}^*, \mathbf{c}, N, A, \gamma$ using the following identities:

$$\begin{aligned} p(\mathbf{u}|\omega, \boldsymbol{\theta}^*, \mathbf{c}, N, A, \gamma) &\propto p(\mathbf{u}, N|\omega, \boldsymbol{\theta}^*, \mathbf{c}) = p(\mathbf{u}, N|\boldsymbol{\theta}^*, \mathbf{c}) = \prod_{i=1}^{N(T)} p(u_i, W_i|c_i, \theta^\star_{c_i}) \\ &= \prod_{i=1}^{N(T)} h_{\theta^\star_{c_i}}(W_i) \frac{1}{\xi_{c_i}} \mathbb{1}_{[0,\xi_{c_i}]}(u_i) \propto \prod_{i=1}^{N(T)} \frac{1}{\xi_{c_i}} \mathbb{1}_{[0,\xi_{c_i}]}(u_i) \end{aligned}$$

where $\xi_{c_i} = \min(w_{c_i}, \zeta)$. So for every $i = 1 \dots N(T)$, $u_i \sim \mathcal{U}_{[0, \min(w_{c_i}, \zeta)]}$.

[1*.b] Secondly, we sample the weights of the empty classes $(w_k)_{k \geq K_{N(T)}+1} | N, \omega, \mathbf{c}, \boldsymbol{\theta}^*, A, \gamma$. Without the slice sampler strategy, there is an infinite number to sample. But, thanks to the slice sampling, we only need to sample a finite number $K^\star$. The weights of the empty classes follow the prior distribution (stick breaking). For $k > K_{N(T)}$,

$$\begin{aligned} v_k &\sim \mathcal{B}(1, A) \\ w_k &= v_k r_{k-1} \\ r_k &= r_{k-1}(1 - v_k) \end{aligned}$$

As explained in Fall and Barat (2012), we know that we have to represent all the components such that their weights $w_k > u_i$ for all the u_i. Assume that we have sampled $w_1, \dots, w_k$, then the weights of the following components can not exceed the rest $1 - \sum_{j=1}^k w_j = r_k$. So if r_k is such that $r_k < u_i$, for all $i = 1 \dots n$, i.e. if $r_k < u^\star$ with $u^\star = \min\{u_1, \dots, u_{N(T)}\}$ we are sure that there is no "interesting component" after that, "interesting" meaning that they won't appear in joint the distribution of $(N, \mathbf{u})$. We can stop and get

$$K^\star = \min\{k, r_k < u^\star\}.$$

In the end we have sampled $(w_{K_{N(T)}+1}, \dots, w_{K^\star})$.

[1*.c] Sample the parameters of the empty classes, $(\theta^*_{K_{N(T)}+1}, \dots, \theta^*_{K^\star})$

$$\forall k = K_{N(T)} + 1, \dots, K^\star, \quad \theta^*_k \sim_{i.i.d} G_\gamma$$

[1*.d] Sample the index $\mathbf{c} = (c_1, \dots, c_{N(T)}) | N, \mathbf{u}, \boldsymbol{\theta}^*, \omega$, i.e. affect the observations to the classes $\{1, \dots, K^\star\}$. Note that we will get new empty classes. The affectations are done using the following probabilities:

$$\begin{aligned} p(\mathbf{c}|\boldsymbol{\theta}^*, \omega, \mathbf{u}, N) &\propto p(\mathbf{c}, \boldsymbol{\theta}^*, \omega, \mathbf{u}, N) = p(N, \mathbf{u}, \mathbf{c}|\boldsymbol{\theta}^*, \omega) p(\omega) p(\boldsymbol{\theta}^*) \\ &\propto \prod_{i=1}^{N(T)} p(W_i, u_i, c_i|\boldsymbol{\theta}^*, \omega) \end{aligned}$$

So, the c_i are independent and, for $i = 1 \dots N(T)$, for $k = 1 \dots, K^\star$

$$p(c_i = k|\boldsymbol{\theta}^*, \omega, u_i, W_i) = w_{i,k} \propto h_{\theta_k^*}(W_i)\frac{w_k}{\min(\zeta, w_k)}\mathbb{1}_{\{k|u_i<\min(\zeta,w_k)\}}(k)$$

We obtain a new $K_{N(T)}$, which is the number of non-empty classes. We re-arrange the weights and the parameters by order of appearance in this affectation.

[$1^\star$.e] Update $(w_1, \dots, w_{K_{N(T)}})$ and $(\theta_1, \dots, \theta_{K_{N(T)}})$ for the non-empty classes.

$$\begin{aligned} p(\theta_k|\mathbf{u}, N, \omega, \mathbf{c}) &\propto G_\gamma(\theta_k) \prod_{i=1, c_i=k}^{n} h_{\theta_k}(W_i), \quad \forall k = 1 \dots K_{N(T)} \quad (1.2) \\ w_1, \dots, w_{K_{N(T)}}, r_{K_{N(T)}} &\sim \text{Dir}(n_1, \dots, n_{K_{N(T)}}, A) \end{aligned}$$

where $n_k = \#\{i|c_i = k\}$ and $r_{K_{N(T)}} = 1 - \sum_{k=1}^{K_{N(T)}} w_k$.
Note that when G_γ is the inverse of the translated inverse exponential distribution, $p(\theta_k|\mathbf{u}, N, \omega, \mathbf{c})$ given in equation (1.2) is:

$$p(\theta_k|\mathbf{u}, N, \omega, \mathbf{c}) \propto \frac{1}{\left(\frac{1}{\theta_k} - \frac{1}{T}\right)^{a-1}} e^{-\frac{\gamma}{\frac{1}{\theta_k} - \frac{1}{T}}} \frac{1}{\theta_k^{n_k}} \mathbb{1}_{\left[\frac{1}{\max_{i|c_i=k} W_i}, +\infty\right[}(\theta_k) \quad (1.3)$$

Its simulation can not be performed directly and we resort to an accept-reject procedure to simulate exactly under this distribution.

Remark 2 *The accept-reject procedure we propose is detailed and discussed here after. Note that (θ_k) changes of dimension at each iteration of the algorithm. As a consequence, a Metropolis-Hastings procedure can not be considered easily, since it could jeopardize the theoretical and practical convergence properties of the algorithm.*

[$2^\star$.] Sampling from $A|\omega, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{u}, \gamma, N$

Let $K_{N(T)}$ be the current number of non-empty classes. West (1992) proves that under the prior distribution $A \sim \Gamma(a_\alpha, b_\alpha)$, we have

$$A|x, K_{N(T)} \sim \pi_x \Gamma(a_A + K_{N(T)}, b_A - \log(x)) + (1 - \pi_x)\Gamma(a_A + K_{N(T)} - 1, b_A - \log(x)) \quad (1.4)$$

where

$$\begin{aligned} x|A, K_{N(T)} &\sim \mathcal{B}(A + 1, N(T)) \\ \frac{\pi_x}{1 - \pi_x} &= \frac{a_A + K_{N(T)} - 1}{n(b_A - \log(x))} \end{aligned}$$

Note that the generation of the new value of A relies on the current value of A.

[$3^\star$.] **Sampling from** $\gamma|\omega, \boldsymbol{\theta}^*, \mathbf{c}, \mathbf{u}, A, N$ If a hierarchical level is set on γ, we have to sample

$$\gamma|\boldsymbol{\theta}^*, \omega, \mathbf{u}, N, \mathbf{c}$$

Using the previous conditional distributions, we have:

$$\begin{aligned} p(\gamma|\boldsymbol{\theta}^*, \omega, \mathbf{u}, N, \mathbf{c}) &\propto p(\mathbf{u}, N, \boldsymbol{\theta}^*, \omega, \mathbf{c}|\gamma)\pi(\gamma) = p(\mathbf{u}, N|\boldsymbol{\theta}^*, \mathbf{c})p(\mathbf{c}|\omega)p(\boldsymbol{\theta}^*|\gamma)\pi(\gamma) = \pi(\gamma)\prod_{k=1}^{K^\star} G_\gamma(\theta_k) \\ &\propto \gamma^{a_\gamma - 1} e^{-b_\gamma \gamma} \prod_{k=1}^{K^\star} \gamma^a e^{-\gamma/(\frac{1}{\theta_k^\star} - \frac{1}{T})} = \gamma^{a_\gamma + aK^\star - 1} e^{-\gamma\left(b_\gamma + \sum_{k=1}^{K^\star} \frac{1}{(\frac{1}{\theta_k^\star} - \frac{1}{T})}\right)} \end{aligned}$$

where $K^\star$ is the total number of classes used to represent $\overline{\lambda}$. Finally, we get:

$$\gamma|\boldsymbol{\theta}^*, \omega, \mathbf{u}, N, \mathbf{c} \sim \Gamma\left(a_\gamma + aK^\star, b_\gamma + \sum_{k=1}^{K^\star} \frac{1}{(\frac{1}{\theta_k^\star} - \frac{1}{T})}\right) \tag{1.5}$$

References

Fall, M. D. and Barat, É. (2012). Gibbs sampling methods for Pitman-Yor mixture models. MAP5 2012-30.

West, M. (1992). Hyperparameter estimation in Dirichlet Process Mixture models. Technical report, Duke University.