Marginal likelihood estimation for Dirichlet Process Mixture models

Adrien Hairault

Joint work with Christian P. Robert and Judith Rousseau
CEREMADE, Université Paris-Dauphine PSL

ISBA@CIRM, June 2021

DPM model

The distribution of a collection of independent observations y $= \{y_1, \dots, y_n\}$ can be modelised by a DPM model as

$$y_i \sim g(y_i|\theta_{s_i}),\ i = 1, \dots, n$$

$$p(s_i = k) = \pi_k,\ i = 1, \dots, n,\ k = 1, 2, \dots$$

$$\pi_1, \pi_2, \dots \sim GEM(\alpha)$$

$$\theta_1, \theta_2, \dots \sim G_0, i.i.d$$

$$\alpha \sim \pi(\alpha)$$

where the *GEM* distribution can be defined through its stick-breaking representation (Sethuraman [1994])

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

for $\{v_i\}_i$ an infinite sequence following the $Beta(1, \alpha)$ distribution ensuring that $\sum_{i=1}^{\infty} w_k = 1$ a.s.

Marginal likelihood for DPM : motivations

Marginal likelihood

$$Z = \int_{\Theta} L(\theta) d\pi(\theta)$$

.

- Many MC algorithms for parametric models : SMC (Del Moral et al. [2006]), Chib's algorithm (Chib [1995]), Nested Samplling (Skilling et al. [2006])...
- Few counterparts for non-parametric models and, to the best of our knowledge, no mainstream algorithm.

Marginal likelihood for DPM : motivations

Why should you care about the marginal likelihood of a non-parametric model ?

- Bayes factor for comparing two models : $B_{01} = \frac{Z_0}{Z_1}$ (Model selection)
- Goodness-of-fit test : comparing a finite mixture model and an 'infinite' mixture model ('the rest of the world')
- Using the DPM carelessly can be dangerous (Miller and Harrison [2014])

A theoretical result for location mixtures (ongoing work with Judith Rousseau and Christian Robert)

Assume $x_1, \ldots, x_n \in \mathbb{R}^d$ arise from a non-degenerate finite location mixture

$$f^*(x) = \sum_{i=1}^{k^*} \pi_i^* \phi_{\Sigma_0}(x - \theta_i^*),\ x \in \mathbb{R}^d,\ \Sigma_0 \in \mathcal{M}^+_{d \times d}(\mathbb{R})$$

Then, under a $DP(\alpha, G_0)$-mixture model where $G_0 = \mathcal{N}_d(\mu, \Sigma)$, the marginal likelihood $m_{DP}(\mathrm{x})$ is such that

$$P^*(m_{DP}(\mathrm{x}) > \eta n^{\frac{-D}{2}}) \longrightarrow 0$$

for $\eta > 0,\ D = (d+1)k^* - 1$

- Implies consistence of the Bayes Factor
- *Still trying to prove a similar result for the location-scale case*

Existing MC algorithms to approximate the evidence in a DPM model

- Basu and Chib [2003] adapt Chib [1995]'s algorithm to the DPM.
  - Rao-Blackwell estimator of the posterior
  - Estimator of the likelihood through SIS is necessary
- Griffin [2017] applies the general SMC framework to the DPM but does not compare its results to Chib
- To our knowledge, those algorithms have never been compared

Reverse Logistic Regression, Geyer [1994]

Let

$$X_1, X_2, X_3, \ldots, X_{n_1} \sim \pi_1$$

and

$$Y_1, Y_2, Y_3, \ldots, Y_{n_2} \sim \frac{\tilde{\pi}_2}{c_2}$$

where $\pi_1, \tilde{\pi}_2$ are known and c_2 unknown.

- Classification problem (logistic regression where the labels are known)
- $\widehat{\log c_2}$ is the intercept of the regression with $\log \pi_1(x), \log \tilde{\pi}_2(x), \log \frac{n_i}{n}$ as the regressors.

Experimental setting

- `galaxy` dataset
- Kernel $g(y|\theta) = \mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2)$
- Conjugate base measure $G_0 = \mathcal{N} - \Gamma^{-1}(\mu_0, \nu, \alpha, \beta)$ so we can integrate θ out and work on the partition induced by the DP
- Prior $\Gamma(1, 1)$ on concentration parameter α
- We use five algorithms : arithmetic mean, harmonic mean, adaptive SMC, Chib's algorithm (Basu and Chib [2003]), and an adaptation of the reverse logistic regression (Geyer [1994])

Results for 6 points from the galaxy dataset

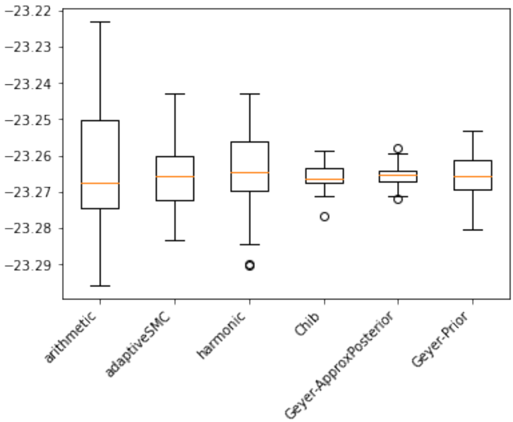

Figure 1: Log marginal likelihood for the different algorithms for $n = 6$, 50 replications each

Results for the full galaxy dataset

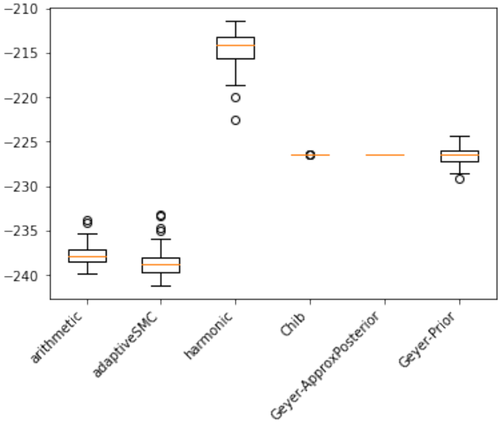

Figure 2: Log marginal likelihood for different algorithms, 40 replications each

SMC convergence

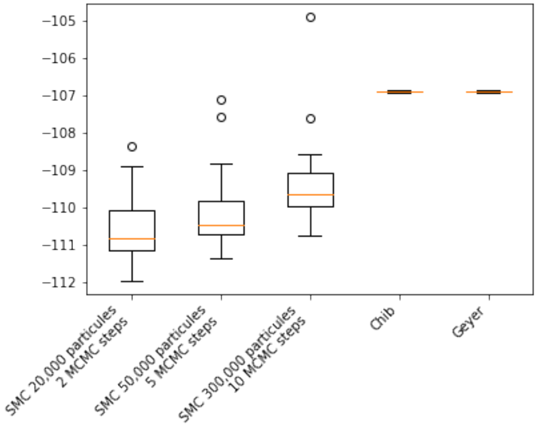

Figure 3: Dataset : 36 points from `galaxy`

Discussion

- Complexity of DPM grows with n and SMC seems to suffer from pathological variance for large n. Need to increase the number of particles and/or the number of MCMC step (or find a better mutation kernel ?) : work in progress, very computationally demanding.
- Though not very popular, the updated algorithm suggested by Basu and Chib [2003] seems to be efficient
- To our knowledge, Geyer [1994] was never applied to the DPM although it seems to perform very well. Maybe easier to use than Chib in the non-conjugate case...

Bibliography I

Sanjib Basu and Siddhartha Chib. Marginal likelihood and bayes factors for dirichlet process mixture models. *Journal of the American Statistical Association*, 98(461):224–235, 2003.

Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the american statistical association*, 90(432): 1313–1321, 1995.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Charles J Geyer. Estimating normalizing constants and reweighting mixtures. 1994.

J.E. Griffin. Sequential monte carlo methods for normalized random measure with independent increments mixtures. *Statistics and Computing*, 27:131–145, 2017.

Bibliography II

Jeffrey W Miller and Matthew T Harrison. Inconsistency of pitman-yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

John Skilling et al. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.