A Practical Implementation of the Gibbs Sampler for Mixture of Distributions: Application to the Determination of Specifications in Food Industry

Julien Cornebise[1], Myriam Maumy[2], and Philippe Girard[3]

[1] E.S.I-E-A
72 avenue Maurice Thorez,
94200 IVRY SUR SEINE, France
(e-mail: cornebis@et.esiea.fr)

LSTA
Université Pierre et Marie Curie - Paris VI
175 rue du Chevaleret,
75013 PARIS, France

[2] IRMA
Université Louis Pasteur
7 rue René Descartes,
67084 STRASBOURG Cedex, France
(e-mail: mmaumy@math.u-strasbg.fr)

[3] Nestlé, Quality Management Department
Av. Nestlé, 55. CH-1800 VEVEY, Switzerland
(e-mail: philippe.girard@nestle.com)

Abstract. This article, mainly targeted to practitioners, illustrates practical issues that may arise when applying MCMC technics to a mixture of distributions model on real data. This data is provided by coffee manufacturer to determine specifications for soluble coffee. Assuming a known number of components, parameters of each component are estimated using the Gibbs sampler and specifications are derived as the 99% quantile of the first distribution. Convergence and label-switching are discussed. Determination of the number of components is also considered, via model selection using the Bayes Factors.

Keywords: MCMC, Mixture, Gibbs Sampler, Label switching, Bayes factors.

1 The Problem and its Modelling

Following an international agreement, a commercial product sold as pure soluble coffee must have been manufactured using green coffee only. However, in a minority of cases, economic adulteration of soluble coffee has been observed in some countries. As a matter of fact, few commercial soluble coffees have been shown to be adulterated with coffee husks/parchments, cereals, and some other plant extracts. In such cases, glucose and xylose contents have proven to be the most discriminant indicators to detect the adulteration. For pure soluble coffee, their concentration are low whereas they become high in case of adulteration. Provided a set of 1002 soluble coffee samples, on which both glucose and xylose concentrations have been measured, we are interested in determining:

- the number K of kinds of production, and their parameters (mean, standard deviation) : $(K-1)$ different frauds, plus one for pure coffee;
- the proportion of each population;
- from the first population and its corresponding characteristics, the specifications within which a soluble coffee can be considered as pure coffee ?

In this article, we only consider the univariate case. Therefore, glucose and xylose concentrations are considered as separate quantities. The approach could, in a further work, be generalised to the bivariate case.

In the univariate case, the observed distribution of the glucose (resp. the xylose) measured on the 1002 coffee samples is modeled as a mixture of normal distributions. We consider that the $T = 1002$ observations from the sample come from K distinct populations (1 pure and $(K-1)$ adulterated), each population $k \in \{1, \ldots, K\}$ following a normal distribution of density f_k and of parameters $\theta_k = (\mu_k, \sigma_k^2)$.

Therefore, the likelihood of an observation x_i, $1 \leq i \leq T$ is:

$$[x_i|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k),$$

where $f_k(\bullet|\theta_k)$ is the probability density function (*pdf*) of a normal distribution with parameters θ_k and π_k is the probability of belonging to population k, such that $\sum_{k=1}^{K} \pi_k = 1$. The parameters of interest are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. The choice of the value of K will be mentioned in section 6.

Other parametrisations for mixtures of normal distributions have been published: interested readers can refer to Robert in [Droesbeke *et al.*, 2002], or in [Marin *et al.*, to appear]. However, this parametrization lacks the natural physical interpretation of the parameters achieved with the actual one.

In the bayesian paradigm, parameters θ_k of each distribution are considered as random variables, having their own distribution. Starting from an initial knowledge about a phenomena described in the *prior distribution* of the parameters $(\boldsymbol{\theta}, \boldsymbol{\pi})$, the Bayes formula enables to update this information by adding the information brought by the data provided the model definition. The *prior* distribution, and its parameters, called *hyperparameters*, are a way to take mathematically into account prior knowledge of the experts of the field, if available (f.i., the potential informations held by the chemists).

To ease the reading, we use throughout the article the notation [.] introduced by [Gelfand *et al.*, 1990] to denote any pdf. In this notation, $[\boldsymbol{\theta}, \boldsymbol{\pi}]$ denotes the *prior* distribution for $(\boldsymbol{\theta}, \boldsymbol{\pi})$, $[y|\boldsymbol{\theta}, \boldsymbol{\pi}]$ the likelihood and $[\boldsymbol{\theta}|\boldsymbol{\pi}, x]$ is the conditional pdf of $\boldsymbol{\theta}$.

Finally, the eventual goal of this application is to estimate a function of the parameters $F(\boldsymbol{\theta}, \boldsymbol{\pi})$ where F can be either the identity function for each parameter or a quantile function. This is generally assessed by

$\mathbb{E}[F(\boldsymbol{\theta},\boldsymbol{\pi})|\boldsymbol{x}] = \int F(\boldsymbol{\theta},\boldsymbol{\pi})[\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{x}]d(\boldsymbol{\theta},\boldsymbol{\pi})$, where $[\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{x}]$ is the pdf of the *posterior* distribution, i.e. the distribution of the parameters conditionnaly to the observations $\boldsymbol{x} = (x_1,\ldots,x_T)$. This pdf is computed via the Bayes formula : $[\boldsymbol{\theta},\boldsymbol{\pi}|\boldsymbol{x}] = [\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\pi}][\boldsymbol{\theta},\boldsymbol{\pi}]/[\boldsymbol{x}]$, where $[\boldsymbol{x}]$ is the *prior* distribution of the observations, which can be taken as constant and thus ignored.

The first question is therefore to choose the *prior* distribution of the parameters, $[\boldsymbol{\theta},\boldsymbol{\pi}]$.

Before going any further, we have to mention that a hidden variable $z_i,\ i \in \{1,\ldots,T\}$ has been introduced in the model mentioned above to ease its Bayesian analysis (see below). This variable is not observed and thus named *latent variable.* $z_i \in \{1,\ldots,K\}$ indicates the original population of the observation x_i, and $\boldsymbol{z} = (z_1,\ldots,z_T)$. The z_i are i.i.d, with $[z_i = k|\boldsymbol{\pi}] \ = \ \pi_k$, $[x_i|\boldsymbol{\theta},\boldsymbol{\pi},z_i = k] = \mathcal{N}(x|\mu_k,\sigma_k^2)$, where $\mathcal{N}(\bullet|\mu,\sigma^2)$ denotes the univariate normal pdf. Analysis of mixture distributions by MCMC methods have been the subject of many publications, for example [Diebolt and Robert, 1990], [Richardson and Green, 1997], [Stephens, 1997], [Marin *et al.*, to appear].

2 Choosing the Prior and its Hyperparameters

As part of the Bayesian analysis, *prior* definition is the first step to go through. Several cases may arise:

- either the experts of the field have valuable information about the distribution of the parameters that should be taken into account : for example, they approximately know what the mean of each component should be.
- or they do not have any information at all - or this information should be ignored, in order to check their results by an objective analysis. Two approaches can be chosen by the statistician :
  - using empirical *prior*, i.e. hyperparameters built upon the data.
  - using non-informative *prior*, i.e. *prior* carrying no information at all. This is somewhat hard to achieve, because purely non-informative *prior* can be improper (for example, uniform distribution on the whole space) and cause troubles. We can also use poorly informative *prior*, for example very dispersed normal distributions.

Moreover, the *prior* is often chosen in a closed-by-sampling or *conjugate prior* family, *i.e.* such that conditionning by the sample (passing from the *prior* to the posterior distribution) only result in a change of the hyperparameters, not in a change of family. This simplifies implementation.

Here we have chosen the following model, that often arises, because each distribution belongs to a closed-by-sampling family :

$$\begin{aligned}\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K) &\text{ sim } \mathrm{Di}(a_1,\ldots,a_K)\\ \mu_k|\sigma_k^2 &\text{ sim } \mathcal{N}(m_k,\sigma_k^2/c_k)\\ \sigma_k^{-2} &\text{ sim } \mathcal{IG}(\alpha_k,\beta_k)\end{aligned}$$

where Di is a Dirichlet distribution, and $\mathcal{IG}$ an Inverse Gamma. Thus, the hyperparameters are a_k, m_k, c_k, α_k, and β_k, $k \in \{1, \ldots, K\}$. For non-informative *prior*, we should have a Uniform distribution for $\boldsymbol{\pi}$, which can be obtained with $a_1 = \ldots = a_K = 1$, a Dirac pdf for σ_k^2, which can be obtained with limit values for α_k and β_k, and a unary density on $\mathbb{R}$ for each μ_k, which is more difficult to implement with limit values on c_k.

In practice, non-informative *prior* is not so easy to deal with : for instance, when programming the algorithm with Matlab, it is not always possible to deal with infinite values of the parameters, of with such particular densities. Therefore, poorly informative *prior*, or even empirical *prior*, should be used. This is what we have done here.

We have chosen : $\forall\, k \in \{1, \ldots, K\}$, $\pi_k = 1$, $c_k = 1$, $m_k = \bar{x}$ (empirical mean), $\alpha_k = K$, and $\beta_k = (T(K-1))^{-1} \sum_{i=1}^{T} (x_i - \bar{x})^2$. Thus, the proportions of each component are non-informative, the means of the μ_k are equal to the empirical mean of the sample, and the means of σ_k^2 is equal to the empirical variance ($\mathbb{E}[\sigma_k^2] = (\alpha_k - 1)\beta_k$ for Inverse Gamma). For reasons that should become clear further (related to label-switching problems and Bayes factors), we have chosen the same hyperparameters for each components: this maintains the symmetry of the density (and therefore of the likelihood).

This part of the Bayesian analysis is certainly the most subjective. The choice of *prior* is clearly the weakest point of the analysis, and the more arguable and argued. Many discussions exist about it, and each approach has its pros and cons. The approach chosen here is neither the most rigorous one, nor the purest, but allows easy implementation. In order to overcome these discussions, a sensitivity analysis needs to be done after the study. Further discussions about the choice of the *prior* can be found in almost any reference: see for example [Droesbeke *et al.*, 2002], [Gelman *et al.*, 2003].

3 Gibbs Sampling, Complete Conditionnal distributions

The evaluation of the expectation is hard to achieve, either analitically or numerically (due either to its complex expression or to its highly multidimensional nature). MCMC methods are a good way to solve this problem. We recall that the principle of Monte-Carlo methods is to generate N independent realizations $(\boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})$ of random variables following the posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{x}]$, and to approximate : $\int F(\boldsymbol{\theta}, \boldsymbol{\pi})[\boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{x}] d(\boldsymbol{\theta}, \boldsymbol{\pi}) \approx \sum_{i=1}^{N} F(\boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})/N$. Here the function F is either the identity function to estimate the parameters or the 99%-quantile function of the first component of the mixture (i.e. the component corresponding to pure coffee powder, without any kind of fraud). In this last case, in order to be the most conservative possible, the empirical 95%-quantile of the sampled values $F(\boldsymbol{\theta}^{(i)}, \boldsymbol{\pi}^{(i)})$, instead of their empirical mean, has been used.

Given the conditional structure of the model of interest, the Gibbs sampler has been used to generate a N-sample from the *posterior* distribution. Quite straightforward to implement, it relies on the availability of all *complete conditional distributions*. Let $\theta = (\theta_1, \dots, \theta_n)$ the vector of the parameters of the model. Here, we have $\theta = (\boldsymbol{\theta}, \boldsymbol{\pi}) = (\mu_1, \sigma_1^2, \dots, \mu_K, \sigma_K^2, \pi_1, \dots, \pi_K)$. Let $\theta_{(i)} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n)$ the vector θ without its i^{th} component. The density of the complete conditional distribution of θ_i is $[\theta_i|\theta_{(i)}, \boldsymbol{x}]$. Let us assume that we know its closed form for all $i \in \{1, \dots, n\}$, wich is often the case. Let us also take arbitrary initial values $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_n^{(0)})$. The Gibbs sampler's algorithm consists in successively sampling:

- $\theta_1^{(i)}$ from $\left[\theta_1|\theta_2^{(i-1)}, \theta_3^{(i-1)}, \theta_4^{(i-1)}, \dots, \theta_n^{(i-1)}\right]$
- $\theta_2^{(i)}$ from $\left[\theta_2|\theta_1^{(i)}\quad, \theta_3^{(i-1)}, \theta_4^{(i-1)}, \dots, \theta_n^{(i-1)}\right]$
- $\cdots$
- $\theta_n^{(i)}$ from $\left[\theta_n|\theta_1^{(i)}\quad, \theta_2^{(i)}\quad, \theta_3^{(i)}\quad, \dots, \theta_{n-1}^{(i)}\right]$

for $i \in \{1, \dots, N + M\}$, where M is a number of iterations that will be discarded. They are called "burn-in" iterations, and correspond to the time before convergence. It can be shown that $\boldsymbol{\theta}^{(M)} = (\theta_1^{(M)}, \dots, \theta_n^{(M)})$ converges in distribution to the posterior joint distribution $[\theta_1, \dots, \theta_n|\boldsymbol{x}]$. The following N values are then considered as a sample from this distribution, and can be used to approximate $F(\theta)$ by empirical mean, as mentioned above.

In our model with the above assumptions, we have the following complete conditional distributions (in order to simplify the notations, the list of all parameters but θ are figured by (θ)):

$$\begin{aligned}
z_i|(\boldsymbol{z_i}), \boldsymbol{x} &\text{ sim } \mathrm{Mu}(\pi_1, \dots \pi_K) \\
\boldsymbol{\pi}|(\boldsymbol{\pi}), \boldsymbol{x} &\text{ sim } \mathrm{Di}(a_1 + n_1, \dots, a_K + n_K), \text{ where } n_k = \textstyle\sum_{i:z_i=k} 1 \\
\mu_k|(\mu_k), \boldsymbol{x} &\text{ sim } \mathcal{N}\left(\tfrac{m_k c_k + s_k}{c_k + n_k}, \tfrac{\sigma_k^2}{n_k + c_k}\right) \\
\sigma_k^{-2}|(\sigma_k^2), \boldsymbol{x} &\text{ sim } \mathcal{IG}\left(\alpha_k + \tfrac{n_k+1}{2}, \beta_k + \tfrac{1}{2}\left(c_k(\mu_k - m_k)^2 + \textstyle\sum_{i:z_i=k}(x_k - \mu_k)^2\right)\right)
\end{aligned}$$

We actually run the sampler with $M = 5000$ and $N = 5000$. Convergence issues and choice of M are discussed in the next section. Metropolis Hastings algorithm details, as well as variants of the Gibbs Sampler can be found in [Droesbeke *et al.*, 2002], or in [Richardson and Green, 1997] or also in [Marin *et al.*, to appear] for an approach more directly linked to mixture of distributions.

4 Convergence Issues

Since the first historically convergence diagnosis (known as the "thick pen" one, [Gelfand *et al.*, 1990]), there exist two main kinds of methods to determine whether the sampler has reached convergence or not, i.e. whether the values from the current generation can be considered as a sample of the target distribution. Two kinds of diagnoses can be considered depending on the number of chains run for carrying out the diagnosis: we consider either the single or multi-chain, where the single ones are rejected by [Gelman and Rubin, 1992b]. We then consider only the multi-chain procedure.

The principle of muti-chains diagnoses is to run multiple independent chains from very different starting points, and to test whether the last values of each chain come from the same distribution or not. If that is the case, we can assume that convergence has been reached. [Gelman and Rubin, 1992a] and [Gelman and Rubin, 1992b], have proposed a method which is often used, based on the comparison of within and between-chains variances. At the beginning of the sampling, the chains are much influenced by the starting point, and the between-chains variance is high above the within-chain one. When each chain has reached the target distribution, the ratio between within and between-chain variance should be around 1. This method has been much improved since then, and some more subtel tests are avalaible, though not implemented here.

If we note $x_{i,j}$ the i^{th} value of chain j, $i \in \{1, \ldots, M+N\}$, $j \in \{1, \ldots, J\}$, we compute the empirical within-chain and between-chain (respectively W and B) as follows

$$\begin{array}{l} W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - x_{\cdot,j}\right)^2 \\ B = \frac{n}{m-1}\sum_{j=1}^{m}\left(x_{\cdot,j} - x_{\cdot,\cdot}\right)^2 \end{array} \text{ with } \begin{array}{l} x_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \\ x_{\cdot,\cdot} = \frac{1}{m}\sum_{j=1}^{m} x_{\cdot,j} \end{array}.$$

The quantity $\hat{\sigma}_+^2 = \frac{n-1}{n}W + \frac{1}{n}B$ can be interpreted as an estimate of the variance of the target distribution. Gelman and Rubin show that, when the initial values of the J chains are chosen "sufficiently different", $\hat{\sigma}_+^2$ systematically overestimates the variance while chains have not reached convergence. Convergence diagnosis is thus based on the statistic $\sqrt{\hat{R}_{GR}} = \sqrt{\hat{\sigma}_+^2/W}$ which tends to 1 when $n \to +\infty$. Practically, convergence is considered as achieved when $\sqrt{\hat{R}_{GR}} < 1,2$. In a multiple parameters case, this diagnosis must be carried out for each parameter separately, the overall convergence being attained when all parameters have converged to its target distribution.

This method, quite straightforward to implement, has proven to be efficient in many cases. However, due to the label-switching issue (see below), this method appeared to be inefficient in our case. Future developments of this study will see this point worked through.

5 The Label Switching Problem

A particularity of the mixture of distributions is that the likelihood and the joint *posterior* pdf (which is the target distribution of the Gibbs Sampler) is symmetrical, i.e. invariant by permutation of the components. Therefore, this last pdf has up to $K!$ duplications of each mode, and the sampler can move from one mode to another freely, thus permuting the components.

The absence of label-switching means that the sampler is stuck in a local mode, maybe because modes are well separated (e.g. when K is very low). The space of parameters is thus not completely explored, which is dangerous.

When estimating a function $F(\boldsymbol{\theta}, \boldsymbol{\pi})$ which is also invariant by permutation of the components (e.g. estimating the pdf at a given point), label-switching is not a problem. But when trying to estimate the parameters of each component, this label-switching has to be undone ([Stephens, 2000b]). Two approaches can be foreseen: imposing an identification constraint during the sampling, or post-processing the generated N-sample to undo the permutation.

The first solution consists of constraining the exploration of the space of parameters by the sampler, which alters the results, see [Celeux *et al.*, 2000], [Marin *et al.*, to appear], [Stephens, 2000b]. Forcing the *prior* to be highly separable (using much different hyperparameters for components) is not a good idea neither : label-switching arises anyway, and the resulting lack of correspondance between the *prior* and the component would corrupt any further use of the *prior* (such as Bayes factor).

We applied here the second solution, i.e. the post-processing. Ordering on μ_k, or on σ_k^2, or on π_k is not a good idea. Some components may be close in mean but not in variance, and vice-versa. Some methods use the Kullback-Leibler "distance" and clustering algorithms (e.g. K-means with $K!$ classes) to determine to which mode (i.e. permutation) belongs each of the sampled vectors of parameters. The reader is invited to read the three references above for more details.

It has to be noted that label-switching is incompatible with convergence diagnoses mentionned in section 4 : comparing the variance between chains is meaningless when components can swap ! Moreover, clustering assumes that convergence has been reached, it would thus be non-sense to use variance-based diagnoses on post-processed samples.

6 Choosing a model : Bayes Factors

Until now, we have worked with a given number K of components. The question is now to choose between different models. Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ be a finite ensemble of possible models (each one with k components, up to K, $K = 7$ in our application) to explain the observations x_i, parametrized by $\boldsymbol{\theta}$. The best model within the K possible is the one with the highest *posterior* probability.

The *posterior* probability of model M_k is calculated via Bayes formula as follows: $[M_k|\boldsymbol{x}] = ([M_k][\boldsymbol{x}|M_k]) \,/\, \left(\sum_{M_k \in \mathcal{M}} [M_k][\boldsymbol{x}|M_k]\right)$, where $[M_k]$ is an *prior* probability of M_k, with $\sum_{M_k \in \mathcal{M}} [M_k] = 1$ (e.g. $\forall\, k$, $[M_k] =$

$1/K$), and $[\boldsymbol{x}|M_k]$, defined by $[\boldsymbol{x}|M_k] = \int[\boldsymbol{x}|\theta_k][\theta_k]\,d\theta_k$ is the *prior* predictive distribution of $\boldsymbol{x}$ under model M_k.

Let us consider M_1 and M_2. The ratio between *posterior* probabilities $[M_2|\boldsymbol{x}]/[M_1|\boldsymbol{x}] = ([M_2][\boldsymbol{x}|M_2])\,/\,([M_1][\boldsymbol{x}|M_1])$ is a *posterior* bet in favor of M_2 compared to M_1. The ratio $B_{21} = [\boldsymbol{x}|M_2]/[\boldsymbol{x}|M_1]$, modifying the *prior* bet $[M_2]/[M_1]$ is called *Bayes factor* of model M_2 relatively to model M_1.

Bayes factors are the basis of bayesian model selection. A ratio close to 1 means that both models equivalently explain the observations, whereas much higher than 1 indicate that the model in the numerator is preferable. [Kass and Raftery, 1995] suggest a scale based on $2\ln(B_{21})$, which can be valid for a first indication, but is far from being general.

The evaluation of Bayes factors, and specially of $[\boldsymbol{x}] = [\boldsymbol{x}|M_k] = \int[\boldsymbol{x}|\theta_k] \times [\theta_k]\,d\theta_k$, relies on the Gibbs outputs. $[\boldsymbol{x}]$ could be estimated by Monte-Carlo sum on the likelihood with values sampled from the *prior* distribution of θ. Nevertheless, *prior* distributions are often very flat, much more than the *posterior* ones, and such method would not be much significant. Newton and Raftery advises to use another method, based on samples from the *posterior* distribution $[\theta|\boldsymbol{x}]$. Bayes formula gives: $[\theta]/[\boldsymbol{x}] = [\theta|\boldsymbol{x}]/[\boldsymbol{x}|\theta]$. By integration on θ, we have: $[\boldsymbol{x}]^{-1} = \int([\theta|\boldsymbol{x}]/[\boldsymbol{x}|\theta])\,d\theta$, and thus can estimate $[\boldsymbol{x}]^{-1}$ by Monte-Carlo methods, sampling from $[\theta|\boldsymbol{x}]$, or more precisely continuing the Gibbs Sampler, setting each parameter one after the other to its estimate, as described in [Chib, 1995] and [Carlin and Chib, 1995].

Again, label-switching is a problem: the estimation by Monte-Carlo method involves here a function which is not invariant by permutation of the component, so permutations have to be undone. That's why, in such cases, other methods such as reversible jump or birth-death processes are preferred (see [Stephens, 2000a]).

7 Possible Improvements, Future Work

This study is an illustration of practical issues encountered when applying MCMC technics for the Bayesian Analysis of the mixture of distribution model. Some issues have already addressed in the literature (label-switching) but without clear and straightforward solutions and some others are pending (*prior* definition). Even though, the methods presented here are quite straightforward to implement, and thus can be easily used in a first approach of the problem.

The following conclusions were attained: due to the recurrent problem of label-switching (caused by the intrisic structure of the dataset), immediate interpretation and efficient model selection (we can not actually choose between 3, 4, or 5 components) were not carried out. Further work in terms of model selection must be definitely done. EM algorithm (using Mixmod software, developed by INRIA's IS2 team) has been used but gave completely different results, incoherent with chemists' interpretations: the algorithm seems trapped in local optimum. MCMC methods are therefore concluded more satisfactory: they at least give meaningfull results.

The following points have to be worked further: the hypothesis of normality of the components (log-normality would seem more coherent), the choice of the *prior* (less informative *prior*, maybe with a learning sub-sample in order to create more informative *prior* that can be used in Bayes factor). A sensitivity analysis is therefore needed to further validate the approach and assess the influence of each assumptions. The question of the convergence is a tricky point to address, provided the label-swistching issue. No rigorous diagnosis has been envisaged: birth-death processes or reversible jump methods need to be considered.

This is an illustration of the difficulties that praticians may face before benefiting from the powerful tools tha are MCMC Bayesian methods.

8 Acknowledgments

The authors would like to thank Eric Parent, from ENGREF (Paris, France), and Luc Perrault, from Hydro-Québec (Varennes, Québec, Canada) for their help and support.

References

[Carlin and Chib, 1995]B.P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *J. R. Statist. Soc. B*, 57:473–484, 1995.

[Celeux *et al.*, 2000]G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.*, 95:957–970, 2000.

[Chib, 1995]S. Chib. Marginal likelihood from the gibbs output. *J. Am. Stat. Assoc.*, 90:1313–1321, 1995.

[Diebolt and Robert, 1990]J. Diebolt and C.P. Robert. Bayesian estimation of finite mixture distributions, part i : Theoretical aspects. Technical Report 110, LSTA, Université Paris VI, 1990.

[Droesbeke *et al.*, 2002]J.J. Droesbeke, J. Fine, and G. Saporta, editors. *Méthodes Bayésiennes en statistique.* Technip, 2002.

[Gelfand *et al.*, 1990]A.E. Gelfand, S.E. Hills, A. Rancine-Poon, and A.F.M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *J. Amer. Stat. Assoc*, 85:972–985, 1990.

[Gelman and Rubin, 1992a]A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequence (with discussion). *Statistical Science*, 7:457–511, 1992.

[Gelman and Rubin, 1992b]A. Gelman and D.B. Rubin. A single series from the gibbs sampler provides a false sense of security (with discussion). In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 625–631. Oxford University Press, 1992.

[Gelman *et al.*, 2003]A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, 2003.

[Kass and Raftery, 1995]R.E. Kass and A.E. Raftery. Bayes factor. *J. Am. Stat. Assoc.*, 90:773–795, 1995.

[Marin *et al.*, to appear]J.M. Marin, K. Mengersen, and C.P. Robert. *Bayesian modelling and inference on mixtures of distributions.* Elsevier-Sciences, (to appear).

[Richardson and Green, 1997]S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B*, 59(4):731–792, 1997.

[Stephens, 1997]M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions.* PhD thesis, Magdalen College, Oxford, 1997.

[Stephens, 2000a]M. Stephens. Bayesian analysis of mixtures with an unknown number of components — an alternative to reversible jump methods. *Annals of Statistics*, 2000.

[Stephens, 2000b]M. Stephens. Dealing with label-switching in mixture models. *J. R. Stat. Soc. B*, 62:795–809, 2000.