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

Julien Cornebise\textsuperscript{1}  Myriam Maumy\textsuperscript{2}  Philippe Girard\textsuperscript{3}

\textsuperscript{1}Ecole Supérieure d’Informatique-Electronique-Automatique, and LSTA, Université Pierre et Marie Curie - Paris VI

\textsuperscript{2}IRMA
Université Louis Pasteur - Strasbourg I

\textsuperscript{3}Quality Management Department, Nestlé

ASMDA 2005, May 18\textsuperscript{th}, 2005
1 The Problem and its Modelling
   - The Coffee Problem
   - Mixture of Normal Laws
   - Bayesian Inference

2 The Use for Markov-Chain Monte-Carlo Methods
   - Choice of the Prior
   - Gibbs Sampler
   - Convergence Checking
   - Label-Switching
   - Model Selection

3 Computer Implementation and Future Work
1 The Problem and its Modelling
   - The Coffee Problem
   - Mixture of Normal Laws
   - Bayesian Inference

2 The Use for Markov-Chain Monte-Carlo Methods
   - Choice of the Prior
   - Gibbs Sampler
   - Convergence Checking
   - Label-Switching
   - Model Selection

3 Computer Implementation and Future Work
The Problem and its Modelling
- The Coffee Problem
- Mixture of Normal Laws
- Bayesian Inference

The Use for Markov-Chain Monte-Carlo Methods
- Choice of the Prior
- Gibbs Sampler
- Convergence Checking
- Label-Switching
- Model Selection

Computer Implementation and Future Work
Outline

1. The Problem and its Modelling
   - The Coffee Problem
   - Mixture of Normal Laws
   - Bayesian Inference

2. The Use for Markov-Chain Monte-Carlo Methods
   - Choice of the Prior
   - Gibbs Sampler
   - Convergence Checking
   - Label-Switching
   - Model Selection

3. Computer Implementation and Future Work
The Problem and its Modelling
The Use for Markov-Chain Monte-Carlo Methods
Computer Implementation and Future Work

The Coffee Problem
Mixture of Normal Laws
Bayesian Inference

---

Pure coffee
1. Manufactured with green coffee only
2. Low glucose rate
3. Low xylose rate

Adulterated coffee
1. Addition of:
   - Husk/Parchment
   - Cereals
   - Other plant extracts...
2. Glucose rate raises
3. Xylose rate raises
Data and Quantities of Interest

Provided:
- a set of 1002 coffee samples’ glucose and xylose rates

Determine:
1. Number $K$ of kinds of production: $(K - 1)$ different frauds, plus one for pure coffee
2. their parameters (mean, standard deviation)
3. their proportions
4. the specifications within which a soluble coffee can be considered as pure coffee
Visualisation of the data

- Histogram of Glucose total for 1002 observations
- Histogram of Xylose total for 1002 observations
Population \( k = 1, \ldots, K \)
- Normally distributed, parameters \( \mu_k, \sigma_k \)
- \( \forall 1 \leq i \leq T = 1002 \), observation \( x_i \) comes from population \( k \) with probability \( \pi_k \), \( \sum_{k=1}^{K} \pi_k = 1 \).

Density of the observations
\[
[x_i | \mu, \sigma, \pi] = \sum_{k=1}^{K} \pi_k N(x_i | \mu_k, \sigma_k), \quad 1 \leq i \leq T = 1002
\]
- \( \mu = (\mu_1, \ldots, \mu_K) \), \( \sigma = (\sigma_i, \ldots, \sigma_K) \), \( \pi = (\pi_1, \ldots, \pi_K) \)
- \( [\cdot|\cdot] \) denotes conditional probability density function (pdf) and parameters of the pdf (bayesian notation, Gelfand et al., 1990).
The Problem and its Modelling

The Use for Markov-Chain Monte-Carlo Methods

Computer Implementation and Future Work

The Coffee Problem

Mixture of Normal Laws

Bayesian Inference

Simple example case, 2 populations

\[ x_i | \pi, \mu, \sigma = \pi \mathcal{N}(x_i | \mu_1, \sigma_1) + (1 - \pi) \mathcal{N}(x_i | \mu_2, \sigma_2) \]

Multiple different shapes:

- \( \mu_1 = -3, \sigma_1 = 1, \mu_2 = 3, \sigma_2 = 1, \pi = 0.5 \)
- \( \mu_1 = -1, \sigma_1 = 4, \mu_2 = 3, \sigma_2 = 1, \pi = 0.4 \)
- \( \mu_1 = 0, \sigma_1 = 4, \mu_2 = 0, \sigma_2 = 1, \pi = 0.5 \)
- \( \mu_1 = -1, \sigma_1 = 4, \mu_2 = 1, \sigma_2 = 1, \pi = 0.8 \)
Augmented data: addition of $z = (z_1, \ldots, z_T)$ to the model, where $z_i$ indicates the population from which observation $x_i$ comes from:

$$\forall i = 1, \ldots, T, \quad z_i \in \{1, \ldots, K\}$$

and

$$[z_i = k] = \pi_k$$

Thus:

$$[x_i | \mu, \sigma, \pi, z] = \mathcal{N}(x_i | \mu_{z_i}, \sigma_{z_i})$$

Other models exist, with many advantages, but lack the immediate physical interpretation (see for example Robert, in Droesbeke et al. (eds), 2002, or Marin et al., in Dey and Rao (eds), to appear in 2005)
Interested in estimating $F(\mu, \sigma, \pi)$, where:

**Function F can be**
- The identity function, to estimate each parameter
- 99%-quantile of the “pure” population
- any other function

Estimated through expectancy of **posterior distribution** $[\mu, \sigma, \pi|\mathbf{x}]$:

**Estimation**

\[
F(\widehat{\mu}, \widehat{\sigma}, \widehat{\pi}) = \mathbb{E}[F(\mu, \sigma, \pi)|\mathbf{x}]
\]

\[
= \int_{\Theta} F(\mu, \sigma, \pi)[\mu, \sigma, \pi|\mathbf{x}]d(\mu, \sigma, \pi)
\]

where $\Theta$ is the space of the parameters, dimension $3K - 1$. 
The posterior density, key of the Bayesian inference, is simply obtained via:

\[
[\mu, \sigma, \pi | x] = \frac{[x | \mu, \sigma, \pi] \times [\mu, \sigma, \pi]}{[x]}
\]

where

- \([x | \mu, \sigma, \pi]\) comes from the model
- \([\mu, \sigma, \pi]\) is the prior distribution, carrying all information available “a priori” (former experiences, experts’ knowledge, etc)
- \([x]\) can be seen as a constant
Outline

1. The Problem and its Modelling
   - The Coffee Problem
   - Mixture of Normal Laws
   - Bayesian Inference

2. The Use for Markov-Chain Monte-Carlo Methods
   - Choice of the Prior
   - Gibbs Sampler
   - Convergence Checking
   - Label-Switching
   - Model Selection

3. Computer Implementation and Future Work
Analysis of mixture of distributions using MCMC methods has been the subject of many publications, for example:

- Diebolt and Robert, 1990,
- Richardson and Green, 1997,
- Stephens, 1997,

Gibbs sampler and connected questions also has been treated in much details, for example by:

- Gelfand et al., 1990
- Gelman and Rubin, 1992
- Carlin and Chib, 1995
- Kass and Raftery, 1995
- Celeux et al., 2000
- Gelman et al., 2003
- ...
On the prior depend the posterior and the complete conditional laws.

Choice of the prior is the most arguable part of Bayesian Analysis ⇒ need for sensitivity analysis.

Two possible cases:
1. Experts have valuable “a priori” information about the parameters, leading to informative prior.
2. No information available, or none to take into account: empirical prior, built on the data, non-informative prior, difficult to really reach, depend on which function of which parameters, improper, or possibility of poorly informative prior.

Hyperparameters are the parameters of the prior distribution.

Conjugate prior is such that going from prior to posterior distribution only results in an update of the parameters: the family of distribution is closed by sampling. Simplifies implementation.
Our choice

We choose to compare two different (conjugate) priors, mentionned respectively (for example) in Marin et al., to appear, and Stephens, 1997.

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi \sim \text{Di}(a_1, \ldots, a_K) )</td>
<td>( \pi \sim \text{Di}(a_1, \ldots, a_K) )</td>
</tr>
<tr>
<td>( \mu_k</td>
<td>\sigma_k^2 \sim N(m_k, \sigma_k^2 / c_k) )</td>
</tr>
<tr>
<td>( \sigma_k^2 \sim IG(\alpha_k, \beta_k) )</td>
<td>( \sigma_k^{-2}</td>
</tr>
<tr>
<td>( \beta \sim \Gamma(g, h) )</td>
<td></td>
</tr>
</tbody>
</table>

where \( k = 1, \ldots, K \), Di is a Dirichlet distribution, \( \Gamma \) a Gamma, and \( IG \) an Inverse Gamma.

**Main difference**: Distribution of the means of the components does or does not depend on variances of the components.
Reason for the need of MCMC Methods

The sum
\[ \int_{\Theta} F(\mu, \sigma, \pi)[\mu, \sigma, \pi|x]d(\mu, \sigma, \pi) \]
is most often intractable, either analytically or numerically, due to either
- its high-dimensional nature
- the complexity of the closed form of the posterior distribution \([\mu, \sigma, \pi|x]\)
- or even the absence of closed form!
Markov-Chain Monte-Carlo Methods’ Principles (1)

“Monte-Carlo” part:

Key Principle

Sample an arbitrary $N$ realisations

$$
\{(\mu^{(j)}, \sigma^{(j)}, \pi^{(j)}): j = 1, \ldots, N\} \text{ from the posterior distribution, }
$$

$$
[\mu, \sigma, \pi|x], \text{ and approximate the expectancy by the average}
$$

$$
\mathbb{E}[F(\mu, \sigma, \pi)|x] = \int_{\Theta} F(\mu, \sigma, \pi)[\mu, \sigma, \pi|x]d(\mu, \sigma, \pi)
$$

$$
\approx \frac{1}{N} \sum_{j=1}^{N} F(\mu^{(j)}, \sigma^{(j)}, \pi^{(j)})
$$
“Markov-Chain” part:
The question now is “How to sample from the posterior distribution”?

Answer: Gibbs Sampler

Build a continuous-state space Markov-Chain on the space of parameters admitting the posterior distribution as its stationary and limit distribution.

This is the purpose of the Gibbs Sampler. More general algorithms exist (such as Metropolis-Hastings), but Gibbs Sampler is very straightforward to implement.
GS relies on the complete conditional laws, which often can easily be sampled from. Let $\theta = (\mu, \sigma, \pi)$:

**Gibbs Sampler**

1. Start from an initial value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$.
2. Then sample successively, for $j = 1, \ldots, M + N$ generations:
   - $\theta_1^{(j)}$ from $\begin{bmatrix} \theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \theta_4^{(j-1)}, \ldots, \theta_n^{(j-1)}, x \end{bmatrix}$
   - $\theta_2^{(j)}$ from $\begin{bmatrix} \theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \theta_4^{(j-1)}, \ldots, \theta_n^{(j-1)}, x \end{bmatrix}$
   - $\theta_3^{(j)}$ from $\begin{bmatrix} \theta_3 | \theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)}, \ldots, \theta_n^{(j-1)}, x \end{bmatrix}$
   - $\ldots$
   - $\theta_n^{(j)}$ from $\begin{bmatrix} \theta_n | \theta_1^{(j)}, \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_{n-1}^{(j)}, x \end{bmatrix}$.
Complete conditionnal laws can be easily calculated using hierarchical graphical model summarizing conditionnal independance relations.

It can be shown that the Gibbs Sampler converges toward the posterior distribution (see e.g. Stephens, 1997, for a demonstration).

The first $M$ iterations are “burn-in” iterations before convergence, discarded.

Though not independent samples, it can be shown that the approximation of the expectancy is still valid.
In the “convenient” cases, the convergence of the Gibbs Sampler can be checked, i.e. the number $M$ of burn-in iterations can be determined.

Sample visualisation of convergence:
Convergence diagnosis based on ANOVA methods

- Originally for univariate chains (1 parameter only), if multiple coordinates are present, diagnose separately for each one.
- Multiple chains are run, and empirical within-chain and between-chain variances are compared (Gelman and Rubin, 1992).
- Let $\theta_{i,j}$ the $i^{th}$ value of chain $j$ - in case of univariate chains, eitherway do the diagnosis for each coordinate of the chains:

\[ W = \frac{1}{m \times (n-1)} \sum_{i,j} (\theta_{i,j} - \theta_{..j})^2 \]
\[ B = \frac{n}{m-1} \sum_j (\theta_{..j} - \theta_{..})^2 \]

with $\theta_{..} = \frac{1}{n} \sum_i \theta_{i,j}$ and $\theta_{..j} = \frac{1}{m} \sum_j \theta_{..j}$

- ANOVA theory gives distributions for $W$ and $B$-based statistics, and thus tests for convergence.
These diagnostics are efficient for single-modal posterior distributions.

But ... Next section will show that mixture models’ posterior distribution is heavily multimodal.

Thus, unable to rigorously check convergence. Should use tools to compare multi-modal distributions.
A “sane pain” : Label-Switching

Source of the problem

The mixture model is not identifiable: the density of the observations

\[
[x_i | \mu, \sigma, \pi] = \sum_{k=1}^{K} \pi_k N(x_i | \mu_k, \sigma_k), \quad 1 \leq i \leq T = 1002
\]

is invariant by permutation of the components, i.e. by relabelling. \( \Rightarrow \) Each mode is replicated \( K! \) times (once for each possible labelling).

The posterior distribution is thus also invariant by permutation, as well as the target distribution of the Gibbs Sampler.
Visualisation of the problem

Two parameters switching

The marginal distributions are exactly similar
So, two possibilities:

1. Either the label-switching doesn’t occurs:
   - ergodic mean is efficient,
   - but based on a sampler that is not mixing enough (1st prior),
   - stay trapped in local maximum of the posterior density:
   - bad exploration of the parameters space,
   - and bad estimations!

2. Or the label-switching heavily occurs:
   - complete exploration of the parameters space,
   - but ergodic mean doesn’t mean anything!

Different priors give different mixing.

Note: If the function of interest is invariant by permutation too, there is no problem.
**Bad ideas**

- Imposing identifiability on the priors: constrains exploration
- Forcing identifiability at each step: constrains exploration

**Failing ideas**

Very clearly explained in Celeux et al., 2000:
- Post-processing, ordering on one of the parameters: not always well-separated
Promising ideas

Algorithms from Stephens, 1997, based on “mode hunting”

1. Uses extension of Kullback-Leibler distance for scaled normal densities
2. Iteratively seek permutations of each generation minimizing a given criterion

Computationnaly heavy, but more efficient.

Nevertheless, fails to separate all modes on our data (see below).
Conclusion : need for other algorithms, or even other samplers.
Example of posterior without L-S

Two components out of $K = 4$, first prior, $M = 1000$, $N = 10000$:
Example of posterior with L-S

Two components out of $K = 4$, second prior, $M = 1000$, $N = 10000$:

J. Cornebise, M. Maumy, P. Girard

A Practical Implementation of the Gibbs Sampler . . .
Two components out of $K = 4$, second prior, $M = 1000$, $N = 10000$, undone:
• Until now, $K$ fixed.
• Model selection : wich value of $K$ ?
• Also formulated as : wich model $M_i$ of $\mathcal{M} = \{M_1, \ldots, M_m\}$ maximises posterior model probability $[M_i|x]$ ?
• Based on evaluation of ratios $[M_j|x]/[M_i|x]$.
• Prior distribution on $M_i : [M_i]$ with $\sum_{i=1}^m [M_i] = 1$
• Prior predictive distribution of $x$ under $M_i$ :

$$[x|M_i] = \int [x|\theta_{M_i}] d\theta_{M_i}$$
We have the posterior bet:

\[
\frac{[M_j|x]}{[M_i|x]} = \frac{[M_j]}{[M_i]} \times \frac{x|M_j}{x|M_i}
\]

**Bayes Factor**

The ratio \( B_{ji} = \frac{x|M_j}{x|M_i} \) modifies the prior bet into a posterior bet. It is called Bayes Factor of model \( M_j \) relatively to model \( M_i \).

- Kass and Raftery, 1995, suggest a scale based on \( 2 \log(B_{ji}) \).
- Evaluation of \( [x|M_i] = \int [x|\theta_{M_i}] d\theta_{M_i} : \text{MCMC too!} \)
- But ... Label-Switching occurs and avoid estimation.
- Possible solution: use other sampler.
Outline

1. The Problem and its Modelling
   - The Coffee Problem
   - Mixture of Normal Laws
   - Bayesian Inference

2. The Use for Markov-Chain Monte-Carlo Methods
   - Choice of the Prior
   - Gibbs Sampler
   - Convergence Checking
   - Label-Switching
   - Model Selection

3. Computer Implementation and Future Work
These computational aspects, though not detailed much here, should not be neglected: with the first prior, $K = 4$, label-switching does not occur before $N = 100000$, risk to miss it.

- Methods implemented using Matlab.
- Massively optimized source code: use of profiling tools, and vectorization of operations.
- Memory accesses and allocation optimized, so that performances do not collapse when $N$ grows.
- Gibbs Sampler's execution time: around 170 iterations / second, 10 000 iterations / minute!
- Comparison with “hands-on” tools such as WinBUGS.
Much improvements before being satisfied

- Get rid of label-switching
- Thus conduct Bayes Factors
- Lead a sensitivity-analysis
- Hypothesis of normality criticized: log-normality?
- Other samplers may avoid many troubles:
  1. Reversible Jump (Richardson and Green, 1997): variable dimension of the states’ space!
  2. Birth-Death Process (Stephens, 1997)
Thank you for your attention!

Please feel free to make any suggestion or question. Any comment is particularly welcome, now, or later by e-mail:

cornebis@et.esiea.fr
mmaumy@math.u-strasbg.fr
philippe.girard@nestle.com