# Variational Bayesian Approximation method for Classification and Clustering with a mixture of Student-t model

Ali Mohammad-Djafari

Laboratoire des signaux et systèmes (L2S),
CNRS-CentraleSupélec-UNIV PARIS SUD, Gif-sur-Yvette, FRANCE,
djafari@lss.supelec.fr,
WWW home page: http://djafari.free.fr

**Abstract.** Clustering, classification and Pattern Recognition in a set of data are between the most important tasks in statistical researches and in many applications. In this paper, we propose to use a mixture of Student-t distribution model for the data via a hierarchical graphical model and the Bayesian framework to do these tasks. The main advantages of this model is that the model accounts for the uncertainties of variances and covariances and we can use the Variational Bayesian Approximation (VBA) methods to obtain fast algorithms to be able to handle large data sets.

## 1 Introduction

Clustering and classification of a set of data are not trivial problems. In fact, we can consider them as ill-posed inverse problems in which the solutions are not unique. Mixture models are natural ones for classification and clustering [1–8]. The Mixture of Gaussians (MoG) models have been used very extensively [9–11]. In this paper, we propose to use a mixture of Student-t model and a Bayesian framework for these tasks. The main advantages of this model is that the model accounts for the uncertainties of variances and covariances and we can use the Variational Bayesian Approximation (VBA) methods to obtain fast algorithms as well. Even if this model may have been used before [12–26], here we propose a novel unifying presentation for all the steps: training, supervised or semi-supervised classification and clustering (non-supervised). We also use VBA framework and some simplifications to develop fast algorithms to be able to handle big data sets.

## 2 Mixture models for classification and clustering

A mixture model is generally given as:

$$p(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{\Theta}, K) = \sum_{k=1}^{K} a_k \, p_k(\boldsymbol{x}_k|\boldsymbol{\theta}_k), \tag{1}$$

where $K$ is the number of classes, $\boldsymbol{a} = \{a_k, k = 1, \cdots, K\}$ the proportion parameters and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k, k = 1, \cdots, K\}$ all the other parameters of the model. If we assume different classes can be modelled by the same family $p_k(\boldsymbol{x}_k|\boldsymbol{\theta}_k) = p(\boldsymbol{x}_k|\boldsymbol{\theta}_k)$ and introduce a hidden class variable $c_n \in \{1, \cdots, K\}$, then for a given sample $\boldsymbol{x}_n$ in class $k$ we can write:

$$p(\boldsymbol{x}_n|c_n = k, \boldsymbol{\theta}_k) = p(\boldsymbol{x}_n|\boldsymbol{\theta}_k) \tag{2}$$

or

$$p(\boldsymbol{x}_n, c_n = k|a_k, \boldsymbol{\theta}_k, K) = a_k\, p(\boldsymbol{x}_n|\boldsymbol{\theta}_k). \tag{3}$$

The Mixture of Gaussians (MoG) corresponds to the case where $p(\boldsymbol{x}_n|c_n = k, \boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Now, imagine a set of data $\boldsymbol{X} = \{\boldsymbol{x}_n, n = 1, \cdots, N\}$ where each element $\boldsymbol{x}_n$ can be in one of these classes. Then, we can write:

$$p(\boldsymbol{X}_n, c_n = k|\boldsymbol{a}, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, c_n = k|\boldsymbol{a}, \boldsymbol{\theta}). \tag{4}$$

Noting by $\boldsymbol{c} = \{c_n, n = 1, N\}$ with $c_n \in \{1, \cdots, K\}$, $\boldsymbol{a} = \{a_k, k = 1, \cdots, K\}$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k, k = 1, \cdots, K\}$, we have:

$$\begin{aligned} p(\boldsymbol{X}_n, \boldsymbol{c}|\boldsymbol{a}, \boldsymbol{\Theta}, K) &= \textstyle\prod_{n=1}^{N} \prod_{k=1}^{K} p(c_n = k)\, p(\boldsymbol{x}_n|\boldsymbol{\theta}_k) \\ &= \textstyle\prod_{n=1}^{N} \prod_{k=1}^{K} a_k\, p(\boldsymbol{x}_n|\boldsymbol{\theta}_k). \end{aligned} \tag{5}$$

The classification problems can then be summarized as follows:

**Training:**
Given a set of (training) data $\boldsymbol{X}$ and classes $\boldsymbol{c}$, estimate the parameters $\boldsymbol{a}$ and $\boldsymbol{\Theta}$. The classical frequentist method is the Maximum Likelihood (ML) which defines the solution as

$$(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\Theta}}) = \arg \max_{(\boldsymbol{a}, \boldsymbol{\Theta})} \left\{ p(\boldsymbol{X}, \boldsymbol{c}|\boldsymbol{a}, \boldsymbol{\Theta}, K) \right\}. \tag{6}$$

The Bayesian way is to assign priors $p(\boldsymbol{a}|K)$ and $p(\boldsymbol{\Theta}|K) = \prod_{k=1}^{K} p(\boldsymbol{\theta}_k)$, then the joint posterior laws is given by:

$$p(\boldsymbol{a}, \boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{c}, K) = \frac{p(\boldsymbol{X}, \boldsymbol{c}|\boldsymbol{a}, \boldsymbol{\Theta}, K)\, p(\boldsymbol{a}|K)\, p(\boldsymbol{\Theta}|K)}{p(\boldsymbol{X}, \boldsymbol{c}|K)} \tag{7}$$

where

$$p(\boldsymbol{X}, \boldsymbol{c}|K) = \iint p(\boldsymbol{X}, \boldsymbol{c}|\boldsymbol{a}, \boldsymbol{\Theta}|K) p(\boldsymbol{a}|K)\, p(\boldsymbol{\Theta}|K)\, \mathrm{d}\boldsymbol{a}\, \mathrm{d}\boldsymbol{\Theta} \tag{8}$$

from which we can deduce $\widehat{\boldsymbol{a}}$ and $\{\widehat{\boldsymbol{\theta}}_k, k = 1, \cdots, K\}$ either as the Maximum A Posteriori (MAP) or Posterior Mean (PM).

**Supervised classification:**
For a given sample $\boldsymbol{x}_m$ and given the parameters $K$, $\boldsymbol{a}$ and $\boldsymbol{\Theta}$ determine

$$p(c_m = k|\boldsymbol{x}_m, \boldsymbol{a}, \boldsymbol{\Theta}, K) = \frac{p(\boldsymbol{x}_m, c_m = k|\boldsymbol{a}, \boldsymbol{\Theta}, K)}{p(\boldsymbol{x}_m|\boldsymbol{a}, \boldsymbol{\Theta}, K)} \tag{9}$$

where $p(\boldsymbol{x}_m, c_m = k|\boldsymbol{a}, \boldsymbol{\Theta}, K) = a_k p(\boldsymbol{x}_m|\boldsymbol{\theta}_k)$ and

$$p(\boldsymbol{x}_m|\boldsymbol{a}, \boldsymbol{\Theta}, K) = \sum_{k=1}^{K} a_k \, p(\boldsymbol{x}_m|\boldsymbol{\theta}_k) \tag{10}$$

and its best class $k^*$, for example the MAP solution:

$$k^* = \arg\max_k \left\{p(c_m = k|\boldsymbol{x}_m, \boldsymbol{a}, \boldsymbol{\Theta}, K)\right\}. \tag{11}$$

**Semi-supervised classification:**
For a given sample $\boldsymbol{x}_m$ and given the parameters $K$ and $\boldsymbol{\Theta}$, determine the probabilities

$$p(c_m = k|\boldsymbol{x}_m, \boldsymbol{\Theta}, K) = \frac{p(\boldsymbol{x}_m, c_m = k|\boldsymbol{\Theta}, K)}{p(\boldsymbol{x}_m|\boldsymbol{\Theta}, K)} \tag{12}$$

where

$$p(\boldsymbol{x}_m, c_m = k|\boldsymbol{\Theta}, K) = \int p(\boldsymbol{x}_m, c_m = k|\boldsymbol{a}, \boldsymbol{\Theta}, K) p(\boldsymbol{a}|K) \, \mathrm{d}\boldsymbol{a} \tag{13}$$

and

$$p(\boldsymbol{x}_m|\boldsymbol{\Theta}, K) = \sum_{k=1}^{K} p(\boldsymbol{x}_m, c_m = k|\boldsymbol{\Theta}, K) \tag{14}$$

and its best class $k^*$, for example the MAP solution:

$$k^* = \arg\max_k \left\{p(c_m = k|\boldsymbol{x}_m, \boldsymbol{\Theta}, K)\right\}. \tag{15}$$

**Clustering or non-supervised classification:**
Given a set of data $\boldsymbol{X}$, determine $K$ and $\boldsymbol{c}$. When these are determined, we can also determine the characteristics of those classes $\boldsymbol{a}$ and $\boldsymbol{\Theta}$. To do this we need the following relations:

$$p(K = L|\boldsymbol{X}) = \frac{p(\boldsymbol{X}, K = L)}{p(\boldsymbol{X})} = \frac{p(\boldsymbol{X}|K = L)\, p(K = L)}{p(\boldsymbol{X})} \tag{16}$$

and

$$p(\boldsymbol{X}) = \sum_{L=1}^{L_0} p(K = L)\, p(\boldsymbol{X}|K = L), \tag{17}$$

where $L_0$ is the a priori maximum number of classes and

$$p(\boldsymbol{X}|K = L) = \int\int \prod_{n} \prod_{k=1}^{L} a_k p(\boldsymbol{x}_n, c_n = k|\boldsymbol{\theta}_k) p(\boldsymbol{a}|K)\, p(\boldsymbol{\Theta}|K) \, \mathrm{d}\boldsymbol{a} \, \mathrm{d}\boldsymbol{\Theta}. \tag{18}$$

As we will see later the main difficulty is the computation of these two last equations. The Variational Bayesian Approximation technics try to find upper and lower bounds for them.

## 3 Mixture of Student-t model

Let us consider the following representation of the Student-t probability density function (pdf):

$$\mathcal{T}(\boldsymbol{x}|\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \int_0^\infty \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},z^{-1}\boldsymbol{\Sigma})\,\mathcal{G}(z|\frac{\nu}{2},\frac{\nu}{2})\,\mathrm{d}z, \tag{19}$$

where

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\\ &= |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left[-\tfrac{1}{2}\mathrm{Tr}\left\{(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})'\right\}\right],\end{aligned} \tag{20}$$

and

$$\mathcal{G}(z|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}\exp\left[-\beta z\right]. \tag{21}$$

Let us also consider the finite mixture of Student-t model:

$$p(\boldsymbol{x}|\{\nu_k,\boldsymbol{a}_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,k=1,\cdots,K\},K) = \sum_{k=1}^K a_k\,\mathcal{T}(\boldsymbol{x}_n|\nu_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \tag{22}$$

Introducing the hidden variables $z_{nk}$ this model can be written via:

$$p(\boldsymbol{x}_n,c_n=k,z_{nk}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,K) = a_k\,\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,z_{n,k}^{-1}\boldsymbol{\Sigma}_k)\,\mathcal{G}(z_{n,k}|\frac{\nu_k}{2},\frac{\nu_k}{2}). \tag{23}$$

Noting by: $\boldsymbol{Z}=\{z_{nk}\}$, $\boldsymbol{z}_k=\{z_{nk},n=1,\cdots,N\}$, $\boldsymbol{c}=\{c_n,n=1,\cdots,N\}$, $\boldsymbol{\theta}_k=\{\nu_k,\boldsymbol{a}_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}$, $\boldsymbol{\Theta}=\{\boldsymbol{\theta}_k,k=1,\cdots,K\}$ and assigning the priors $p(\boldsymbol{\Theta})=\prod_k p(\boldsymbol{\theta}_k)$, we can write:

$$p(\boldsymbol{X},\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|K) = \prod_n\prod_k a_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,z_{n,k}^{-1}\boldsymbol{\Sigma}_k)\,\mathcal{G}(z_{nk}|\tfrac{\nu_k}{2},\tfrac{\nu_k}{2})\,p(\boldsymbol{\theta}_k) \tag{24}$$

Then, the joint posterior law of all the unknowns $(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta})$ given the data $\boldsymbol{X}$ and $K$ can be written as

$$p(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|\boldsymbol{X},K) = \frac{p(\boldsymbol{X},\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|K)}{p(\boldsymbol{X}|K)}. \tag{25}$$

The main task now is to propose some approximations to it in such a way that we can use it easily in all the above mentioned tasks of classification or clustering. The main idea behind the VBA technics is exactly this.

## 4 Variational Bayesian Approximation (VBA)

### 4.1 Main idea

The main idea behind the VBA is to propose an approximation $q(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta})$ for $p(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|\boldsymbol{X},K)$. This approximation can be in such a way that $\mathrm{KL}(q:p)$ be minimized. Interestingly, by noting that $p(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|\boldsymbol{X},K) = p(\boldsymbol{X},\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|K)/p(\boldsymbol{X}|K)$, it is easy to showed that

$$\mathrm{KL}(q:p) = -\mathcal{F}(q) + \ln p(\boldsymbol{X}|K) \tag{26}$$

where

$$\mathcal{F}(q) = \langle -\ln p(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{Z}, \boldsymbol{\Theta}|K)\rangle_q \tag{27}$$

is called free energy of $q$ and we have the following properties:
– Maximizing $\mathcal{F}(q)$ or minimizing $\mathrm{KL}(q : p)$ are equivalent and both give un upper bound to the evidence of the model $\ln p(\boldsymbol{X}|K)$.
– When the optimum $q^*$ is obtained, $\mathcal{F}(q^*)$ can be used as a criterion for model selection.
– If $p$ is in the exponential family, then choosing appropriate conjugate priors, the structure of $q$ will be the same and we can obtain appropriate fast optimization algorithms.

In our case, noting that

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{Z}, \boldsymbol{\Theta}|K) = &\prod_n \prod_k p(\boldsymbol{x}_n, c_n, z_{nk}|a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) \\ &\prod_k \left[p(\alpha_k)\, p(\beta_k)\, p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)\, p(\boldsymbol{\Sigma}_k)\right] \end{aligned} \tag{28}$$

with

$$p(\boldsymbol{x}_n, c_n, z_{nk}|a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, z_{n,k}^{-1}\boldsymbol{\Sigma}_k)\, \mathcal{G}(z_{nk}|\alpha_k, \beta_k) \tag{29}$$

is separable, in one side for $[\boldsymbol{c}, \boldsymbol{Z}]$ and in other size in components of $\boldsymbol{\Theta}$, we propose to use

$$q(\boldsymbol{c}, \boldsymbol{Z}, \boldsymbol{\Theta}) = q(\boldsymbol{c}, \boldsymbol{Z})\, q(\boldsymbol{\Theta}). \tag{30}$$

With this decomposition, the expression of the Kullback-Leibler divergence becomes:

$$\begin{aligned} &\mathrm{KL}(q_1(\boldsymbol{c}, \boldsymbol{Z})q_2(\boldsymbol{\Theta}) : p(\boldsymbol{c}, \boldsymbol{Z}, \boldsymbol{\Theta}|\boldsymbol{X}, K) = \\ &\sum_{\boldsymbol{c}} \int\int q_1(\boldsymbol{c}, \boldsymbol{Z})q_2(\boldsymbol{\Theta}) \ln \frac{q_1(\boldsymbol{c}, \boldsymbol{Z})q_2(\boldsymbol{\Theta})}{p(\boldsymbol{c}, \boldsymbol{Z}, \boldsymbol{\Theta}|\boldsymbol{X}, K)}\, \mathrm{d}\boldsymbol{\Theta}\, \mathrm{d}\boldsymbol{Z} \end{aligned} \tag{31}$$

and the expression of the Free energy becomes:

$$\mathcal{F}(q_1(\boldsymbol{c}, \boldsymbol{Z})q_2(\boldsymbol{\Theta})) = \sum_{\boldsymbol{c}} \int\int q_1(\boldsymbol{c}, \boldsymbol{Z})q_2(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{Z}|\boldsymbol{\Theta}, K)p(\boldsymbol{\Theta}|K)}{q_1(\boldsymbol{c}, \boldsymbol{Z})q_2(\boldsymbol{\Theta})}\, \mathrm{d}\boldsymbol{\Theta}\, \mathrm{d}\boldsymbol{Z}. \tag{32}$$

In the following we propose appropriate priors and obtain the expressions of $q$ and appropriate fast algorithms.

## 5 Proposed VBA for Mixture of Student-t priors model

As we discussed in previous section, here we consider the Mixture of Student-t priors model and propose appropriate conjugate priors and appropriate factorized form for the testing or approximation $q$ and finally give the details of the parameters updating algorithm. To be able to propose conjugate priors for all the parameters, we change slightly the model by replacing $\nu_k$ in the Gamma expression $\mathcal{G}(z_{n,k}|\frac{\nu_k}{2}, \frac{\nu_k}{2})$ of the Student-t expression by two parameters $\mathcal{G}(z_{n,k}|\alpha_k, \beta_k)$:

$$p(\boldsymbol{x}_n, c_n = k, z_{nk}|a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \beta_k, K) = a_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, z_{n,k}^{-1}\boldsymbol{\Sigma}_k)\, \mathcal{G}(z_{n,k}|\alpha_k, \beta_k). \tag{33}$$

The final hierarchical model that we propose is shown in the Figure 1.

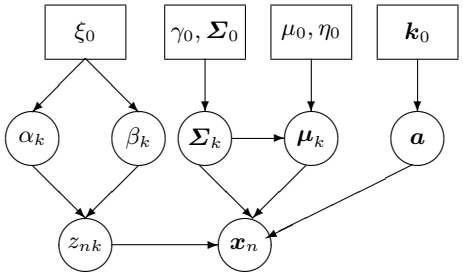


**Fig. 1.** Graphical representation of the model.

### 5.1 Conjugate priors

In the following, noting by $\boldsymbol{\Theta} = \{(a_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \beta_k), k = 1, \cdots, K\}$, we propose to use the factorized prior laws:

$$p(\boldsymbol{\Theta}) = p(\boldsymbol{a}) \sum_k \left[p(\alpha_k)\, p(\beta_k)\, p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)\, p(\boldsymbol{\Sigma}_k)\right] \tag{34}$$

with the following components:

$$\begin{cases} p(\boldsymbol{a}) = \mathcal{D}(\boldsymbol{a}|\boldsymbol{k}_0), \quad \boldsymbol{k}_0 = [k_0, \cdots, k_0] = k_0 1 \\ p(\alpha_k) = \mathcal{E}(\alpha_k|\zeta_0) = \mathcal{G}(\alpha_k|1, \zeta_0) \\ p(\beta_k) = \mathcal{E}(\beta_k|\zeta_0) = \mathcal{G}(\alpha_k|1, \zeta_0) \\ p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mu_0 1, \eta_0^{-1}\boldsymbol{\Sigma}_k) \\ p(\boldsymbol{\Sigma}_k) = \mathcal{IW}(\boldsymbol{\Sigma}_k|\gamma_0, \gamma_0 \Sigma_0) \end{cases} \tag{35}$$

where

$$\mathcal{D}(\boldsymbol{a}|\boldsymbol{k}) = \frac{\Gamma(\sum_l k_k)}{\prod_l \Gamma(k_l)} \prod_l a_l^{k_l - 1} \tag{36}$$

is the Dirichlet pdf,

$$\mathcal{E}(t|\zeta_0) = \zeta_0 \exp\left[-\zeta_0 t\right] \tag{37}$$

is the Exponential pdf,

$$\mathcal{G}(t|a, b) = \frac{b^a}{\Gamma(a)} t^{a-1} \exp\left[-bt\right] \tag{38}$$

is the Gamma pdf and

$$\mathcal{IW}(\boldsymbol{\Sigma}|\gamma,\gamma\boldsymbol{\Delta}) = \frac{|\frac{1}{2}\boldsymbol{\Delta}|^{\gamma/2}\exp\left[-\frac{1}{2}\mathrm{Tr}\left\{\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}\right\}\right]}{\Gamma_D(\gamma/2)|\boldsymbol{\Sigma}|^{\frac{\gamma+D+1}{2}}}. \tag{39}$$

is the inverse Wishart pdf.

With these prior laws and the likelihood: $p(\boldsymbol{x}_n|c(n),\boldsymbol{z}_k(n),\boldsymbol{\Theta},k)$ we can obtain the joint posterior law:

$$p_k(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta})}{p(\boldsymbol{X})}. \tag{40}$$

Now, we have to choose a factored form for $q$ in such a way that we can transform the optimization of the $KL(q:p)$ or the free energy $\mathcal{F}(q)$ to the updating of the parameters of the different components of $q$. We propose to use the following decomposition:

$$\begin{aligned} q(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}) = q(\boldsymbol{c},\boldsymbol{Z})\,q(\boldsymbol{\Theta}) = \\ \prod_n\prod_k[q(c_n=k|z_{nk})\,q(z_{nk})] \\ \prod_k[q(\alpha_k)\,q(\beta_k)\,q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)\,q(\boldsymbol{\Sigma}_k)]\,q(\boldsymbol{a}). \end{aligned} \tag{41}$$

with:

$$\begin{cases} q(\boldsymbol{a}) = \mathcal{D}(\boldsymbol{a}|\widetilde{\boldsymbol{k}}), \quad \widetilde{\boldsymbol{k}} = [\widetilde{k}_1,\cdots,\widetilde{k}_K] \\ q(\alpha_k) = \mathcal{G}(\alpha_k|\widetilde{\zeta}_k,\widetilde{\eta}_k) \\ q(\beta_k) = \mathcal{G}(\beta_k|\widetilde{\zeta}_k,\widetilde{\eta}_k) \\ q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\widetilde{\boldsymbol{\mu}},\widetilde{\eta}^{-1}\boldsymbol{\Sigma}_k) \\ q(\boldsymbol{\Sigma}_k) = \mathcal{IW}(\boldsymbol{\Sigma}_k|\widetilde{\gamma},\widetilde{\gamma}\widetilde{\Sigma}) \end{cases} \tag{42}$$

With these choices, we have

$$\begin{aligned} \mathcal{F}(q(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta})) &= \langle\ln p(\boldsymbol{X},\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta}|K)\rangle_{q(\boldsymbol{c},\boldsymbol{Z},\boldsymbol{\Theta})} \\ &= \prod_k\prod_n\mathcal{F}_{1_{kn}} + \prod_k\mathcal{F}_{2_k} \end{aligned} \tag{43}$$

with

$$\begin{aligned} \mathcal{F}_{1_{kn}} &= \langle\ln p(\boldsymbol{x}_n,c_n,z_{nk},\boldsymbol{\theta}_k)\rangle_{q(c_n=k|z_{nk})q(z_{nk})} \\ \mathcal{F}_{2_k} &= \langle\ln p(\boldsymbol{x}_n,c_n,z_{nk},\boldsymbol{\theta}_k)\rangle_{q(\boldsymbol{\theta}_k)} \end{aligned} \tag{44}$$

Now, to obtain the expressions of the updating expressions of the tilded parameters, we need to go to the following three steps:

- E step: Optimizing $\mathcal{F}$ with respect to $q(\boldsymbol{c},\boldsymbol{Z})$ when keeping $q(\boldsymbol{\Theta})$ fixed, we obtain the expression of $q(c_n=k|z_{nk}) = \widetilde{a}_k$, $q(z_{nk}) = \mathcal{G}(z_{nk}|\widetilde{\alpha}_k,\widetilde{\beta}_k)$.
- M step: Optimizing $\mathcal{F}$ with respect to $q(\boldsymbol{\Theta})$ when keeping $q(\boldsymbol{c},\boldsymbol{Z})$ fixed, we obtain the expression of $q(\boldsymbol{a}) = \mathcal{D}(\boldsymbol{a}|\widetilde{\boldsymbol{k}})$, $\widetilde{\boldsymbol{k}} = [\widetilde{k}_1,\cdots,\widetilde{k}_K]$, $q(\alpha_k) = \mathcal{G}(\alpha_k|\widetilde{\zeta}_k,\widetilde{\eta}_k)$, $q(\beta_k) = \mathcal{G}(\beta_k|\widetilde{\zeta}_k,\widetilde{\eta}_k)$, $q(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\widetilde{\boldsymbol{\mu}},\widetilde{\eta}^{-1}\boldsymbol{\Sigma}_k)$, and $q(\boldsymbol{\Sigma}_k) = \mathcal{IW}(\boldsymbol{\Sigma}_k|\widetilde{\gamma},\widetilde{\gamma}\widetilde{\Sigma})$, which gives the updating algorithm for the corresponding tilded parameters.

– $\mathcal{F}$ evaluation After each E step and M step, we can also evaluate the expression of $\mathcal{F}(q)$ which can be used for stopping rule of the iterative algorithm. Also, final value of this expression for each value of $K$, noted $\mathcal{F}_k$, can be used as a criterion for the model selection, i.e.; the determination of the number of clusters.

The expressions of all the tilded parameters update as well as the expression of $\mathcal{F}_K$ are easily obtained thanks to the properties of the conjugate priors. However, these expressions are cumbersome and will be given in the appendix.

## 6 Conclusion

Clustering and classification of a set of data are between the most important tasks in statistical researches for many applications such as data mining in biology. Mixture models and in particular Mixture of Gaussians are classical models for these tasks. In this paper, we proposed to use a mixture of Student-t distribution model for the data via a hierarchical graphical model. Then, we proposed a Bayesian framework to do these tasks. The main advantages of this model is that the model accounts for the uncertainties of variances and covariances and we can use the Variational Bayesian Approximation (VBA) methods. To obtain fast algorithms and be able to handle large data sets, we used conjugate priors everywhere it was possible. The proposed algorithm has been used for clustering, classification and discriminent analysis of some biological data, but in this paper, we only presented the main algorithm.

## Acknowledgment

This work has been supported partially by the C5SYS (https://www.erasysbio.net/index.php?index=272) project of ERASYSBIO.

## References

1. Forgy, E.W.: Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics (1965)
2. MacKay, D.J.C.: A practical Bayesian framework for backpropagation networks. Neural Computation **4** (1992) 448–472
3. Redner, R., Walker, H.: Mixture densities, maximum likelihood and the em algorithm. SIAM Review **26** (1984)
4. Husmeier, D., Penny, W., Roberts, S.: Empirical evaluation of Bayesian sampling for neural classifiers. In L.Niklason, M., T.Ziemke, eds.: ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks. (1998)
5. Lee, T., Lewicki, M., Sejnowski, T.: Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. Neural Computation **11** (1999) 409–433
6. A., H., E., O.: Independent component analysis: Algorithms and applications. Neural Networks **13** (2000) 411–430

7. Ma, J., Xu, L., Jordan, M.I.: Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. Neural Computation **12** (2001) 2881–2907
8. Nielsen, F., Nock, R.: Clustering multivariate normal distributions. Neural Computation **Springer-Verlag, Berlin, Heidelberg** (2009) 164–174
9. Quandt, R., Ramsey, J.: Estimating mixtures of normal distributions and switching regressions. J. American Statistical Association **73** (1978) 2
10. Hathaway, R.: A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Annals of Statistics **13** (1985) 1
11. Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixture. Journal of Royal Statistical Society **58** (1996) 155–176
12. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. J. R. Statist. Soc. B **39** (1977) 3
13. Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in graphical models **89** (1998) 355–368
14. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal components analysis. Neural Computation **11** (1999) 443–482
15. Jordan, M., Ghahramani, Z., Jaakkola, T., , Saul, L.: An introduction to variational methods for graphical models. Machine Learning **37** (2006) 183–233
16. Jaakkola, T.S., Jordan, M.I.: Bayesian parameter estimation via variational methods. Statistics and Computing **10** (2000) 25–37
17. Friston, K., Penny, W.: Bayesian inference and posterior probability maps. In: Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02). (2002) 413–417
18. Winn, J., Bishop, C.M., Jaakkola, T.: Variational message passing. Journal of Machine Learning Research **6** (2005) 661–694
19. David, M.B., Michael, I.J.: Variational inference for the dirichlet process mixtures. **1** (2006) 121–144
20. Beal, M.: Variational Algorithms for Approximate Bayesian Inference. PhD thesis, Gatsby Computational Neuroscience Unit, University College London (2003)
21. Beal, M., Ghahramani, Z.: Variational Bayesian learning of directed graphical models with hidden variables. Bayesian Statistics **1** (2006) 793–832
22. Kim, H., Ghahramani, Z.: Bayesian gaussian process classification with the em-ep algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 1948–1959
23. Nasios, N., Bors, A.: Variational learning for gaussian mixture models. IEEE Transactions on Systems, Man and Cybernetics, Part B **36** (2006) 849–862
24. Ghahramani, Z., Griffiths, T., Sollich, P.: Bayesian nonparametric latent feature models. Bayesian Statistics **8** (2007)
25. McGrory, C., Titterington, D.: Variational approximations in Bayesian model selection for finite mixture distributions. Computational Statistics and Data Analysis **51** (2007) 5352–5367
26. Qiao, Z., Zhou, L., Huang, J.Z.: Sparse linear discriminant analysis with applications to high dimensional low sample size data. International Journal of Applied Mathematics **39** (2008) 48–60