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

Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes (L2S)
UMR8506 CNRS-CentraleSupélec-UNIV PARIS SUD
SUPELEC, 91192 Gif-sur-Yvette, France
http://lss.centralesupelec.fr
Email: djafari@lss.supelec.fr
http://djafari.free.fr
http://publicationslist.org/djafari

Contents

Mixture models

- General mixture model

$$p(\mathbf{x}|\mathbf{a},\mathbf{\Theta},K)=\sum_{k=1}^{K} a_k\, p_k(\mathbf{x}_k|\boldsymbol{\theta}_k),\quad 0<a_k<1$$

- Same family $p_k(\mathbf{x}_k|\boldsymbol{\theta}_k)=p(\mathbf{x}_k|\boldsymbol{\theta}_k),\ \forall k$
- Gaussian $p(\mathbf{x}_k|\boldsymbol{\theta}_k)=\mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ with $\boldsymbol{\theta}_k=(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$
- Data $\mathbf{X}=\{\mathbf{x}_n,n=1,\cdots,N\}$ where each element $\mathbf{x}_n$ can be in one of these classes c_n.
- $a_k=p(c_n=k)$, $\mathbf{a}=\{a_k,k=1,\cdots,K\}$, $\mathbf{\Theta}=\{\boldsymbol{\theta}_k,k=1,\cdots,K\}$

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

Different problems

- Training:
  Given a set of (training) data $\mathbf{X}$ and classes $\mathbf{c}$, estimate the parameters $\mathbf{a}$ and $\mathbf{\Theta}$.
- Supervised classification:
  Given a sample $\mathbf{x}_m$ and the parameters K, $\mathbf{a}$ and $\mathbf{\Theta}$ determine its class

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{a}, \mathbf{\Theta}, K)\}.$$

- Semi-supervised classification (Proportions are not known):
  Given sample $\mathbf{x}_m$ and the parameters K and $\mathbf{\Theta}$, determine its class

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{\Theta}, K)\}.$$

- Clustering or unsupervised classification (Number of classes K is not known):
  Given a set of data $\mathbf{X}$, determine K and $\mathbf{c}$.

Training

- Given a set of (training) data $\mathbf{X}$ and classes $\mathbf{c}$, estimate the parameters $\mathbf{a}$ and $\mathbf{\Theta}$.
- Maximum Likelihood (ML):

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{\Theta}}) = \underset{(\mathbf{a},\mathbf{\Theta})}{\arg\max} \left\{ p(\mathbf{X}, \mathbf{c}|\mathbf{a}, \mathbf{\Theta}, K) \right\}.$$

- Bayesian: Assign priors $p(\mathbf{a}|K)$ and $p(\mathbf{\Theta}|K) = \prod_{k=1}^{K} p(\theta_k)$ and write the expression of the joint posterior laws:

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

where

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

- Infer on $\mathbf{a}$ and $\mathbf{\Theta}$ either as the Maximum A Posteriori (MAP) or Posterior Mean (PM).

Supervised classification

- Given a sample $\mathbf{x}_m$ and the parameters K, $\mathbf{a}$ and $\mathbf{\Theta}$ determine

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

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

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

- Best class k^*:

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{a}, \mathbf{\Theta}, K)\}$$

Semi-supervised classification

- Given sample $\mathbf{x}_m$ and the parameters K and $\mathbf{\Theta}$ (not the proportions $\mathbf{a}$), determine the probabilities

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

where

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

and

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

- Best class k^*, for example the MAP solution:

$$k^* = \arg\max_k \{p(c_m = k|\mathbf{x}_m, \mathbf{\Theta}, K)\}.$$

Clustering or non-supervised classification

- Given a set of data $\mathbf{X}$, determine K and $\mathbf{c}$.
- Determination of the number of classes:

$$p(K = L|\mathbf{X}) = \frac{p(\mathbf{X}, K = L)}{p(\mathbf{X})} = \frac{p(\mathbf{X}|K = L)\, p(K = L)}{p(\mathbf{X})}$$

and

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

where L_0 is the a priori maximum number of classes and

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

- When K and $\mathbf{c}$ are determined, we can also determine the characteristics of those classes $\mathbf{a}$ and $\boldsymbol{\Theta}$.

Mixture of Student-t model

- Student-t and its Infinite Gaussian Scaled Model (IGSM):

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

where

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

and

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

- Mixture of Student-t:

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

Mixture of Student-t model

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

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

- Joint posterior law:

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

- 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.

Variational Bayesian Approximation (VBA)

- Main idea: to propose easy computational approximation $q(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta})$ for $p(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|\mathbf{X}, K)$.
- Criterion: $\mathrm{KL}(q : p)$
- Interestingly, by noting that
  $p(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|\mathbf{X}, K) = p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|K)/p(\mathbf{X}|K)$
  we have:

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

  where

$$\mathcal{F}(q) = \left\langle -\ln p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|K) \right\rangle_q$$

  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 $\ln p(\mathbf{X}|K)$.
  – When the optimum q^* is obtained, $\mathcal{F}(q^*)$ can be used as a criterion for model selection.

VBA: choosing the good families

- Using $\mathrm{KL}(q:p)$ has the very interesting property that using q to compute the means we obtain the same values if we have used p (Conservation of the means).
- Unfortunately, this is not the case for variances or other moments.
- 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.

Hierarchical graphical model

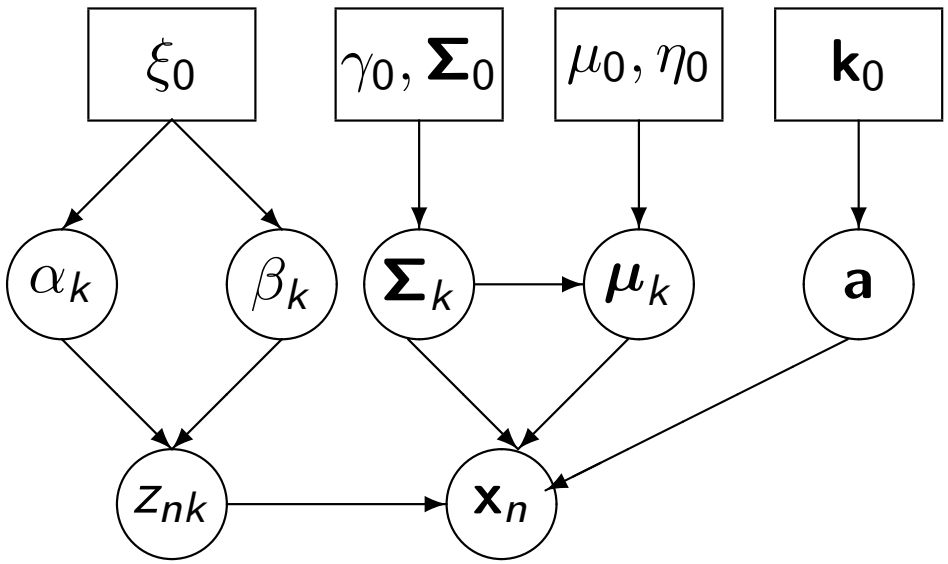

Figure : Graphical representation of the model.

VBA for mixture of Student-t

- In our case, noting that

$$p(\mathbf{X}, \mathbf{c}, \mathbf{Z}, \mathbf{\Theta}|K) = \prod_n \prod_k p(\mathbf{x}_n, c_n, z_{nk}|a_k, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \nu_k) \prod_k [p(\alpha_k)\, p(\beta_k)\, p(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k)\, p(\mathbf{\Sigma}_k)]$$

with

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

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

$$q(\mathbf{c}, \mathbf{Z}, \mathbf{\Theta}) = q(\mathbf{c}, \mathbf{Z})\, q(\mathbf{\Theta}).$$

VBA for mixture of Student-t

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

$$\mathrm{KL}(q_1(\mathbf{c},\mathbf{Z})q_2(\mathbf{\Theta}) : p(\mathbf{c},\mathbf{Z},\mathbf{\Theta}|\mathbf{X},K) = \sum_{\mathbf{c}} \int\int q_1(\mathbf{c},\mathbf{Z})q_2(\mathbf{\Theta}) \ln \frac{q_1(\mathbf{c},\mathbf{Z})q_2(\mathbf{\Theta})}{p(\mathbf{c},\mathbf{Z},\mathbf{\Theta}|\mathbf{X},K)} \,\mathrm{d}\mathbf{\Theta}\,\mathrm{d}\mathbf{Z}$$

- The expression of the Free energy becomes:

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

Proposed VBA for Mixture of Student-t priors model

- Using a generalized Student-t obtained by replacing $\mathcal{G}(z_{n,k}|\frac{\nu_k}{2},\frac{\nu_k}{2})$ by $\mathcal{G}(z_{n,k}|\alpha_k,\beta_k)$ it will be easier to propose conjugate priors for α_k,β_k than for ν_k.

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

- 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:

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

with the following components:

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

Proposed VBA for Mixture of Student-t priors model

where

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

is the Dirichlet pdf,

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

is the Exponential pdf,

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

is the Gamma pdf and

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

is the inverse Wishart pdf.

With these prior laws and the likelihood: joint posterior law:

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

Expressions of q

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

with:

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

With these choices, we have

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

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

VBA Algorithm step

Expressions of the updating expressions of the tilded parameters are obtained by following three steps:

- E step: Optimizing $\mathcal{F}$ with respect to $q(\mathbf{c}, \mathbf{Z})$ when keeping $q(\mathbf{\Theta})$ fixed, we obtain the expression of $q(c_n = k|z_{nk}) = \tilde{a}_k$, $q(z_{nk}) = \mathcal{G}(z_{nk}|\widetilde{\alpha}_k, \widetilde{\beta}_k)$.
- M step: Optimizing $\mathcal{F}$ with respect to $q(\mathbf{\Theta})$ when keeping $q(\mathbf{c}, \mathbf{Z})$ fixed, we obtain the expression of $q(\mathbf{a}) = \mathcal{D}(\mathbf{a}|\tilde{\mathbf{k}})$, $\quad \tilde{\mathbf{k}} = [\tilde{k}_1, \cdots, \tilde{k}_K]$, $q(\alpha_k) = \mathcal{G}(\alpha_k|\tilde{\zeta}_k, \tilde{\eta}_k)$, $q(\beta_k) = \mathcal{G}(\beta_k|\tilde{\zeta}_k, \tilde{\eta}_k)$, $q(\boldsymbol{\mu}_k|\mathbf{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\widetilde{\boldsymbol{\mu}}, \tilde{\eta}^{-1}\mathbf{\Sigma}_k)$, and $q(\mathbf{\Sigma}_k) = \mathcal{IW}(\mathbf{\Sigma}_k|\tilde{\gamma}, \tilde{\gamma}\tilde{\Sigma})$, 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.
- Final value of $\mathcal{F}(q)$ for each value of K, noted $\mathcal{F}_k$, can be used as a criterion for model selection, i.e.; the determination of the number of clusters.

Conclusions

- 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.
- We proposed to use a mixture of generalised Student-t distribution model for the data via a hierarchical graphical model.
- 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 discriminant analysis of some biological data (Cancer research related), but in this paper, we only presented the main algorithm.