Multivariate "Diagonal" Mixture Models

Serge Iovleff

October 22, 2013

Abstract

This document resume the different "diagonal" mixture models that are, or will be, implemented in the stk++ Clustering project and used in the MixtComp project.

1 A Short Course on Mixture Modeling

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be n independent vectors in $\mathbb{R}^d$ such that each $\mathbf{x}_i$ arises from a probability distribution with density

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k, \boldsymbol{\alpha}) \tag{1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, ..., K$ and $p_1 + ... + p_K = 1$), $h(\cdot|\boldsymbol{\lambda}_k, \boldsymbol{\alpha})$ denotes a d-dimensional distribution parameterized by $\boldsymbol{\lambda}_k$ and $\boldsymbol{\alpha}$. The parameters $\boldsymbol{\alpha}$ do not depend from k and is common to all the components of the mixture. The vector parameter to be estimated is $\theta = (p_1, \ldots, p_K, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K, \boldsymbol{\alpha})$ and is chosen to maximize the observed log-likelihood

$$L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} p_k h(\mathbf{x}_i, \boldsymbol{\lambda}_k, \boldsymbol{\alpha}) \right). \tag{2}$$

It is well known that for a mixture distribution, a sample of indicator vectors or *labels* $\mathbf{z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$, $z_{ik} = 1$ or 0, according to the fact that $\mathbf{x}_i$ is arising from the kth mixture component or not, is associated to the observed data $\mathbf{x}$. The sample $\mathbf{z}$ is *unknown* so that the maximum likelihood estimation of mixture models is performed via the EM algorithm [Dempster et al., 1997] or by a stochastic version of EM called SEM (see [Mclachlan and Peel, 2000]), or by a k-means like algorithm called CEM.

1.1 The algorithms

1.1.1 The EM algorithm

Starting from an initial arbitrary parameter θ^0, the mth iteration of the EM algorithm consists of repeating the following E and M steps.

- **E step:** The current conditional probabilities that $z_{ik} = 1$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ are computed using the current value θ^{m-1} of the parameter:

$$t_{ik}^m = t_k^m(\mathbf{x}_i|\theta^{m-1}) = \frac{p_k^{m-1} h(\mathbf{x}_i|\boldsymbol{\lambda}_k^{m-1}, \boldsymbol{\alpha}^{m-1})}{\sum_{l=1}^K p_l^{m-1} h(\mathbf{x}_i|\boldsymbol{\lambda}_l^{m-1}, \boldsymbol{\alpha}^{m-1})}. \tag{3}$$

- **M step:** The m.l. estimate θ^m of θ is updated using the conditional probabilities t_{ik}^m as conditional mixing weights. It leads to maximize

$$L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{t}^m) = \sum_{i=1}^n \sum_{k=1}^K t_{ik}^m \ln\left[p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k, \boldsymbol{\alpha})\right], \tag{4}$$

  where $\mathbf{t}^m = (t_{ik}^m, i = 1, \ldots, n, k = 1, \ldots, K)$. Updated expression of mixture proportions are, for $k = 1, \ldots, K$,

$$p_k^m = \frac{\sum_{i=1}^n t_{ik}^m}{n}. \tag{5}$$

  Detailed formula for the updating of the $\boldsymbol{\lambda}_k$'s and $\boldsymbol{\alpha}$ are depending of the component parameterization and will be detailed in the next section.

The EM algorithm may converge to a local maximum of the observed data likelihood function, depending on starting values.

1.1.2 The SEM algorithm

The SEM algorithm is a stochastic version of EM incorporating between the E and M steps a restoration of the unknown component labels $\mathbf{z}_i$, $i = 1, \ldots, n$, by drawing them at random from their current conditional distribution. Starting from an initial parameter θ^0, an iteration of SEM consists of three steps.

- **E step:** The conditional probabilities t_{ik}^m $(1 \leq i \leq n, 1 \leq k \leq K)$ are computed for the current value of θ as done in the E step of EM.
- **S step:** Generate labels $\mathbf{z}^m = \{\mathbf{z}_1^m, ..., \mathbf{z}_n^m\}$ by assigning each point $\mathbf{x}_i$ at random to one of the mixture components according to the categorical distribution with parameter $(t_{ik}^m, 1 \leq k \leq K)$.
- **M step:** The m.l. estimate of θ is updated using the generated labels by maximizing

$$L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{t}^m) = \sum_{i=1}^n \sum_{k=1}^K z_{ik}^m \ln\left[p_k h(\mathbf{x}_i|\boldsymbol{\lambda}_k, \boldsymbol{\alpha})\right], \tag{6}$$

SEM does not converge point wise. It generates a Markov chain whose stationary distribution is more or less concentrated around the m.l. parameter estimator. A natural parameter estimate from a SEM sequence $(\theta^r)_{r=1,\ldots,R}$ is the mean $\sum_{r=b+1}^R \theta^r/(R-b)$ of the iterates values where the first b burn-in iterates have been discarded when computing this mean. An alternative estimate is to consider the parameter value leading to the highest likelihood in a SEM sequence.

1.1.3 The CEM algorithm

This algorithm incorporates a classification step between the E and M steps of EM. Starting from an initial parameter θ^0, an iteration of CEM consists of three steps.

- **E step:** The conditional probabilities t_{ik}^m $(1 \leq i \leq n, 1 \leq k \leq K)$ are computed for the current value of θ as done in the E step of EM.
- **C step:** Generate labels $\mathbf{z}^m = \{\mathbf{z}_1^m, ..., \mathbf{z}_n^m\}$ by assigning each point $\mathbf{x}_i$ to the component maximizing the conditional probability $(t_{ik}^m, 1 \leq k \leq K)$.
- **M step:** The m.l. estimate of θ are computed as done in the M step of SEM.

CEM is a *K-means*-like algorithm and contrary to EM, it converges in a finite number of iterations. CEM is not maximizing the observed log-likelihood L (2) but is maximizing in θ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ the complete data log-likelihood

$$CL(\theta, \mathbf{z}_1, \ldots, \mathbf{z}_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln[p_k h(\mathbf{x}_i | \boldsymbol{\lambda}_k)]. \tag{7}$$

where the missing component indicator vector $\mathbf{z}_i$ of each sample point is included in the data set. As a consequence, CEM is not expected to converge to the m.l. estimate of θ and yields inconsistent estimates of the parameters especially when the mixture components are overlapping or are in disparate proportions (see [Mclachlan and Peel, 2000], Section 2.21).

2 Multivariate Gaussian Mixture Models

A Gaussian density on $\mathbb{R}_+$ is a density of the form:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \quad \sigma > 0. \tag{8}$$

A joint Gaussian density on ${\mathbb{R}^d}_+$ is a density of the form:

$$h(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{j=1}^{d} f(x^j; \mu^j, \sigma^j) \quad \sigma^j > 0. \tag{9}$$

The parameters $\boldsymbol{\mu} = (\mu^1, \ldots, \mu^d)$ are the position parameters and the parameters $\boldsymbol{\sigma} = (\sigma^1, \ldots, \sigma^d)$ are the standard-deviation parameters. Assumptions on the position and standard-deviation parameters among the variables and the components lead to define four families of mixture model.

Let us write a multidimensional Gaussian mixture model in the from `f_s*` with `b*`, the different ways to parametrize the standard-deviation parameters of a Gaussian mixture:

- `sjk` means that we have one standard-deviation parameter for each variable and for each component,

- `sk` means that the standard-deviation parameters are the same for all the variables inside a component,
- `sj` means that the standard-deviation parameters are different for each variable but are equals between the components,
- and finally `s` means that the standard-deviation parameters are all equals.

The `f_sjk` model is the most general model and have a density function of the form

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} g(x_i^j|\mu_k^j, \sigma_k^j). \tag{10}$$

3 Multivariate Gamma Mixture Models

A gamma density on $\mathbb{R}_+$ is a density of the form:

$$g(x; a, b) = \frac{(x)^{a-1}\, e^{-x/b}}{\Gamma(a)\,(b)^a} \quad a > 0, \quad b > 0. \tag{11}$$

A joint gamma density on ${\mathbb{R}^d}_+$ is a density of the form:

$$h(\mathbf{x}; \mathbf{a}, \mathbf{b}) = \prod_{j=1}^{d} g(x^j; a^j, b^j) \quad a^j > 0, \quad b^j > 0. \tag{12}$$

The parameters $\mathbf{a} = (a^1, \ldots, a^d)$ are the shape parameters and the parameters $\mathbf{b} = (b^1, \ldots, b^d)$ are the scale parameters. Assumptions on the scale and shape parameters among the variables and the components lead to define twelve families of mixture model. Let us write a multidimensional gamma mixture model in the from `g_a*_b*` with `a*` (resp. `b*`), the different ways to parametrize the shape (resp. scale) parameters of a gamma mixture:

- `ajk` (resp. `bjk`) means that we have one shape (resp. scale) parameter for each variable and for each component,
- `ak` (resp. `bk`) means that the shape (resp. scale) parameters are the same for all the variables inside a component,
- `aj` (resp. `bj`) means that the shape (resp. scale) parameters are different for each variable but are equals between the components,
- and finally `a` (resp. `b`) means that the shape (resp. scale) parameters are the same for all the variables and all the components.

The models we can build in this way are summarized in the table 1, in parenthesis we give the number of parameters of each models. The tested and implemented model we present in this article are in the gray cells.

The `g_ajk_bjk` model is the most general and have a density function of the form

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} g(x_i^j|a_k^j, b_k^j). \tag{13}$$

	ajk	ak	aj	a
bjk	g_ajk_bjk (2dK)	g_ak_bjk (dK + K)	g_aj_bjk (dK+d)	g_a_bjk (dK+1)
bk	g_ajk_bk (dK+K)	g_ak_bk (2K)	g_aj_bk (K+d)	g_a_bk (K+1)
bj	g_ajk_bj (dK+d)	g_ak_bj (K+d)	NA	NA
b	g_ajk_b (dK+1)	g_ak_b (K+1)	NA	NA

Table 1: The twelve multidimensional gamma mixture models. In parenthesis the number of parameter of each model. In gray the interesting models.

All the other models can be derived from this model by dropping the indexes in j and/or k from the expression (13). For example the mixture model `g_aj_bk` have a density function of the form

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} g(x_i^j|a^j, b^k). \tag{14}$$

4 Multivariate Beta Mixture Models

A beta density on $(0,1)$ is a density of the form:

$$b(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \quad \alpha > 0, \quad \beta > 0. \tag{15}$$

A joint beta density on $(0,1)^d$ is a density of the form:

$$h(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{j=1}^{d} b(x^j;\alpha^j,\beta^j) \quad \alpha^j > 0, \quad \beta^j > 0. \tag{16}$$

The parameters $\boldsymbol{\alpha} = (\alpha^1, \ldots, \alpha^d)$ and the parameters $\boldsymbol{\beta} = (\beta^1, \ldots, \beta^d)$ are the shape parameters. Assumptions on these parameters among the variables and the components lead to define twelve families of mixture models. Let us write a multidimensional beta mixture model in the form `beta_a*_b*` with `a*` (resp. `b*`). The models we can build are summarized in the table 2 below

	ajk	ak	aj	a
bjk	b_ajk_bjk	b_ak_bjk	b_aj_bjk	b_a_bjk
bk	b_ajk_bk	b_ak_bk	b_aj_bk	b_a_bk
bj	b_ajk_bj	b_ak_bj	NA	NA
b	b_ajk_b	b_ak_b	NA	NA

Table 2: The twelve multidimensional beta mixture models

The `b_ajk_bjk` model is the most general and has a density function of the form

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} b(x_i^j|\alpha_k^j, \beta_k^j). \tag{17}$$

All the other models can be derived from this model by dropping the indexes in j and/or k from the expression (17). For example the mixture model `g_aj_bk` has a density function of the form

$$f(\mathbf{x}_i|\theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} b(x_i^j|\alpha^j, \beta^k). \tag{18}$$

References

[Brent, 1973] Brent, R. P. (1973). Some efficient algorithms for solving systems of nonlinear equations. 10(2):327–344.

[Dempster et al., 1997] Dempster, A., Laird, N., and Rubin, D. (1997). Maximum likelihood from incomplete data with the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1.

[Mclachlan and Peel, 2000] Mclachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1 edition.

A M step computation

In this section, given the array $\mathbf{t}$ of conditional probabilities, we will write $t_{.k} = \sum_{i=1}^{n}$, for $k = 1, \ldots, K$ and we will denote

$$\bar{x}_k^j = \frac{1}{t_{.k}} \sum_{i=1}^{n} t_{ik} x_i^j,$$

the k-th pondered mean of the j-th observation, and by

$$(\overline{\log(x)})_k^j = \frac{1}{t_{.k}} \sum_{i=1}^{n} t_{ik} \log(x_i^j),$$

the k-th pondered log-mean of the j-th observation.

A.1 M Step of the gamma_ajk_bjk model

Replacing h by its expression in the equation (4) and summing in i, we see that we have to maximize in (a_k^j, b_k^j), for $j = 1, \ldots d$ and $k = 1, \ldots, K$

$$l = t_{.k} \left((a_k^j - 1)(\overline{\log(x)})_k^j - \frac{\bar{x}_k^j}{b_k^j} - \log(\Gamma(a_k^j)) - a_k^j \log(b_k^j) \right), \tag{19}$$

which is the weighted version of the usual log-likelihood of the gamma distribution. The maximum likelihood for b_k^j is easily found to be

$$\hat{b}_k^j = \frac{\bar{x}_k^j}{a_k^j}. \tag{20}$$

Deriving in a_k^j and using (20) gives a_k^j solution in a of the following equation

$$(\overline{\log(x)})_k^j - \Psi(a) - \log(\bar{x}_k^j) + \log(a) = 0 \tag{21}$$

whom solution can be found using Brent's method [Brent, 1973].

A.2 M Step of the gamma_ajk_bj model

Replacing h by its expression in the equation (4) and summing in i, we see that we have to maximize in $(a_1^j, \dots, a_K^j, b^j)$, for $j = 1, \dots d$,

$$L = \sum_{k=1}^{K} t_{.k} \left((a_k^j - 1)(\overline{\log(x)})_k^j - \frac{\bar{x}_k^j}{b^j} - \log(\Gamma(a_k^j)) - a_k^j \log(b^j) \right). \tag{22}$$

Deriving in a_k^j and b^j and equaling the derivative to zero, we get the following set of equations

$$\left\{ \begin{array}{lcl} \Psi(a_k^j) & = & (\overline{\log(x)})_k^j - \log(b^j), \quad k = 1, \dots, K \\ b^j & = & \frac{\sum_{k=1}^K t_{.k} \bar{x}_k^j}{\sum_{k=1}^K t_{.k} a_k^j} \end{array} \right. \tag{23}$$

This set of non-linear equation can be solved using an iterative algorithm:

- **Initialization:** Compute an initial estimator of the $\mathbf{a}_k$, $k = 1, \dots K$ and $\mathbf{b}$ using moment estimators.
- **Repeat until convergence :**
  - **a step:** For fixed b^j solve for each a_k^j, the equation:

    $$\Psi(a) - (\overline{\log(x)})_k^j + \log(b^j) = 0.$$

  - **b step:** Update b^j using equation (23).

This algorithm minimize alternatively the log-likelihood in $\mathbf{a}_k$, $k = 1, \dots n$ and in $\mathbf{b}$ and converge in few iterations.

A.3 M Step of the gamma_aj_bjk model

Replacing h by its expression in the equation (4) and summing in i, we see that we have to maximize in $(a^j, \dots, a^j, b_k^j)$, for $j = 1, \dots d$,

$$L = \sum_{k=1}^{K} t_{.k} \left((a^j - 1)(\overline{\log(x)})_k^j - \frac{\bar{x}_k^j}{b_k^j} - \log(\Gamma(a^j)) - a_k^j \log(b_k^j) \right). \tag{24}$$

The maximum likelihood for b_k^j is easily found to be

$$\hat{b}_k^j = \frac{\bar{x}_k^j}{a^j}. \tag{25}$$