Block clustering with mixture model: comparison of different approaches

Mohamed Nadif[1] and Gérard Govaert[2]

[1] LITA - Université de Metz
Ile du Saulcy,
57045 Metz, France
(e-mail: nadif@iut.univ-metz.fr)
[2] Heudiasyc, UMR CNRS 6599
Université de technologie de Compiègne
BP 20529,
60205 Compiègne, France
(e-mail: gerard.govaert@utc.fr)

Abstract. When the data consists of a set of objects described by a set of variables, we have recently proposed a new mixture model which takes into account the block clustering problem on the both sets. In considering this problem under the maximum likelihood and classification maximum likelihood approaches, one can wonder about the performances of the algorithm obtained by block EM, block CEM or by simple uses of the EM and CEM algorithms applied on the both sets separately. The main objective of this paper is to compare these algorithms.
Keywords: Block clustering, Mixture model, EM and CEM algorithms.

1 Introduction

Cluster analysis is an important tool in a variety of scientific areas such as pattern recognition, information retrieval, microarray, data mining, and so forth. Although many clustering procedures such as hierarchical clustering, k-means or self-organizing maps, aim to construct an optimal partition of objects or, sometimes, of variables, there are other methods, called block clustering methods, which consider simultaneously the two sets and organize the data into homogeneous blocks. If $\mathbf{x}$ denotes a $n \times r$ data matrix defined by $\mathbf{x} = \{(x_i^j); i \in I \text{ and } j \in J\}$, where I is a set of n objects (rows, observations, cases) and J is a set of r variables (columns, attributes), the basic idea of these methods consists in making permutations of objects and variables in order to draw a correspondence structure on $I \times J$. These last years, block clustering (also called biclustering) has become an important challenge in data mining context. In the text mining field, [Dhillon, 2001] has proposed a spectral block clustering method by exploiting the clear duality between rows (documents) and columns (words). In the analysis of microarray data where data are often presented as matrices of expression levels of genes under different conditions, block clustering of genes and conditions has permitted

to overcome the problem of the choice of similarity on the both sets found in conventional clustering methods [Cheng and Church, 2000].

The mixture model is undoubtedly one of the greatest contributions to clustering. It offers a great flexibility and solutions to the problem of the number of clusters. To take into account the block clustering situation, we have defined in [Govaert and Nadif, 2003] a block mixture model and, setting the clustering problem in the classification maximum likelihood (CML) approach [Symons, 1981], we have developed an algorithm called block CEM which is based on the alternated application of classical CEM on intermediate data matrices. More recently, setting the clustering problem in the maximum likelihood (ML) approach, we have proposed [Govaert and Nadif, 2005] a generalized EM algorithm (GEM) [Dempster *et al.*, 1977] which maximizes a variational approximation of the likelihood using an iterative algorithm whose steps are carried out by the application of the EM algorithm on intermediate mixture models. In estimation context, we have shown that this approach gives good results on simulated data.

This paper focuses on the clustering context. It deals to compare five algorithms: block CEM, block EM with two variants, two-way EM and two-way CEM, i.e. EM and CEM applied separately on I and J. Results on simulated data are given, confirming that block EM gives much better performance than the other algorithms.

In the following, for convenience, we represent a partition $\mathbf{z}$ into g clusters of the sample I by the vector $(z_1, \ldots, z_n)$, where $z_i \in \{1, \ldots, g\}$ indicates the component of the observation i or by the classification matrix $(z_{ik}, i = 1, \ldots, n, k = 1, \ldots, g)$ where $z_{ik} = 1$ if i belongs to cluster k and 0 otherwise. We will use similar notation for a partition $\mathbf{w}$ into m clusters of the set J. Moreover, to simplify the notation, the sums and the products relating to categories, row clusters will be subscripted respectively by letters i, j and k without indicating the limits of variation which will be thus implicit. Thus, for example, the sum $\sum_i$ stands for $\sum_{i=1}^{n}$ or $\sum_{i,j,k,\ell}$ stands for $\sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{g}\sum_{\ell=1}^{m}$.

2 Block Mixture Model

For the classical mixture model, the probability density function of a mixture sample $\mathbf{x}$ is defined by $f(\mathbf{x};\boldsymbol{\theta}) = \prod_i \sum_k p_k \varphi(\mathbf{x}_i;\boldsymbol{\alpha}_k)$ where the p_k's are the mixing proportions, the $\varphi(\mathbf{x}_i;\boldsymbol{\alpha}_k)$'s are the densities of each component k, and $\boldsymbol{\theta} = (p_1, \ldots, p_g, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$. We have shown [Govaert and Nadif, 2003] that $f(\mathbf{x};\boldsymbol{\theta})$ can be written as

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{z};\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}), \tag{1}$$

where $\mathcal{Z}$ denotes the set of all possible partitions of I in g clusters, $p(\mathbf{z};\boldsymbol{\theta}) = \prod_i p_{z_i}$ and $f(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) = \prod_i \varphi(\mathbf{x}_i;\boldsymbol{\alpha}_{z_i})$. With this formulation, the data matrix $\mathbf{x}$ is assumed to be a sample of size 1 from a random (n, r) matrix.

To study the block clustering problem, we have extended the formulation (1) to propose a block mixture model defined by the following probability density function $f(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{u}\in U} p(\mathbf{u};\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{u};\boldsymbol{\theta})$ where U denotes the set of all possible partitions of $I \times J$ and $\boldsymbol{\theta}$ is the parameter of this mixture model. In restricting this model to a set of partitions of $I \times J$ defined by a product of partitions of I and J, which will be supposed to be independent, we obtain the following decomposition

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z};\boldsymbol{\theta}) p(\mathbf{w};\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta})$$

where $\mathcal{Z}$ and $\mathcal{W}$ denote the sets of all possible partitions $\mathbf{z}$ of I and $\mathbf{w}$ of J.

Now, extending the latent class principle of local independence to our block model, the x_i^j will be supposed to be independent once $\mathbf{z}_i$ and $\mathbf{w}_j$ are fixed; then, we have $f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \prod_{i,j} \varphi(x_i^j;\boldsymbol{\alpha}_{z_i w_j})$ where $\varphi(x,\boldsymbol{\alpha}_{k\ell})$ is a probability density function defined on the real set $\mathbb{R}$. Denoting $\boldsymbol{\theta} = (\mathbf{p},\mathbf{q},\boldsymbol{\alpha}_{11},\ldots,\boldsymbol{\alpha}_{gm})$ where $\mathbf{p} = (p_1,\ldots,p_g)$ and $\mathbf{q} = (q_1,\ldots,q_m)$ are the vectors of probabilities p_k and q_ℓ that a row and a column belong to the kth component and to the ℓth component respectively, we obtain a block mixture model with the following probability density function

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_i p_{z_i} \prod_j q_{w_j} \prod_{i,j} \varphi(x_i^j;\boldsymbol{\alpha}_{z_i w_j}).$$

3 Various approaches

To tackle the block clustering problem, we have used the block mixture model and have considered the ML and CML approaches.

3.1 ML approach and block EM algorithm

For the ML approach, to estimate the parameters of the block mixture model, we proposed to maximize the log-likelihood $L(\boldsymbol{\theta};\mathbf{x}) = \log(f(\mathbf{x};\boldsymbol{\theta}))$ by using the EM algorithm. To describe this algorithm, we must define the complete log-likelihood, also called classification log-likelihood $L_C(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{x},\mathbf{z},\mathbf{w}) = \log f(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})$ which can be written

$$L_C(\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \sum_{i,k} z_{ik} \log p_k + \sum_{j,\ell} w_{j\ell} \log q_\ell + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log \varphi(x_i^j;\boldsymbol{\alpha}_{k\ell}).$$

The EM algorithm maximizes $L(\boldsymbol{\theta};\mathbf{x})$ iteratively by maximizing the conditional expectation of the complete log-likelihood given a previous current estimate $\boldsymbol{\theta}^{(c)}$ and $\mathbf{x}$:

$$\begin{aligned} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(c)}) = &\sum_{i,k} P(z_{ik}=1|\mathbf{x},\boldsymbol{\theta}^{(c)}) \log p_k + \sum_{j,\ell} P(w_{j\ell}=1|\mathbf{x},\boldsymbol{\theta}^{(c)}) \log q_\ell \\ &+ \sum_{i,j,k,\ell} P(z_{ik}w_{j\ell}=1|\mathbf{x},\boldsymbol{\theta}^{(c)}) \log \varphi(x_i^j;\boldsymbol{\alpha}_{k\ell}). \end{aligned}$$

Unfortunately, difficulties arise due to the dependence structure in the model, and more precisely, to the determination of $P(z_{ik}w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)})$ and approximations are required to make the algorithm tractable. Using a variational approximation

$$P(z_{ik}w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)}) \approx P(z_{ik} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)})P(w_{j\ell} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)}),$$

we proposed [Govaert and Nadif, 2005] to maximize alternatively two conditional expectations of the complete-data log-likelihood $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}|\mathbf{d})$ and $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}|\mathbf{c})$ where $\mathbf{c}$ and $\mathbf{d}$ are the matrices defined by the c_{ik}'s and the $d_{j\ell}$'s. We shown that these conditional expectations are associated respectively to classical mixture models

$$\sum_k p_k \psi_k(\mathbf{u}_i; \boldsymbol{\theta}, \mathbf{d}) \text{ and } \sum_\ell q_\ell \psi_\ell(\mathbf{v}^j; \boldsymbol{\theta}, \mathbf{c})$$

where $\mathbf{u}_i = (u_i^1, \ldots, u_i^m)$ and $\mathbf{v}^j = (v_1^j, \ldots, v_g^j)$ are vectors of sufficient statistics and ψ_k and ψ^ℓ are the probability density functions of the sufficient statistics. So, these maximizations can be carried out by the EM algorithm and we obtain the two following versions, called block EM(1) and block EM(2). The different steps of the first one are

1. Start from $\mathbf{c}^{(0)}$, $\mathbf{d}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$.
2. Compute $(\mathbf{c}^{(c+1)}, \mathbf{d}^{(c+1)}, \boldsymbol{\theta}^{(c+1)})$ starting from $(\mathbf{c}^{(c)}, \mathbf{d}^{(c)}, \boldsymbol{\theta}^{(c)})$:
   (a) Compute $\mathbf{c}^{(c+1)}, \mathbf{p}^{(c+1)}, \alpha^{(c+\frac{1}{2})}$ by using on the data $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ the EM algorithm starting from $\mathbf{c}^{(c)}, \mathbf{p}^{(c)}, \alpha^{(c)}$.
   (b) Compute $\mathbf{d}^{(c+1)}, \mathbf{q}^{(c+1)}, \alpha^{(c+1)}$ by using on the data $(\mathbf{v}^1, \ldots, \mathbf{v}^r)$ the EM algorithm starting from $\mathbf{d}^{(c)}, \mathbf{q}^{(c)}, \alpha^{(c+\frac{1}{2})}$.
3. Iterate the steps 2 until convergence.

The different steps of the second version are

1. Start from $\mathbf{c}^{(0)}$, $\mathbf{d}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$, initial values of $\mathbf{c}$, $\mathbf{d}$ and $\boldsymbol{\theta}$.
2. Compute $(\mathbf{c}^{(c+1)}, \mathbf{d}^{(c+1)})$ starting from $\boldsymbol{\theta}^{(c)}$ by iterating the following two steps (a) and (b) until convergence:
   (a) Compute $\mathbf{c}^{(c+1)}$ by using on the data $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ the E-step starting from $\mathbf{d}^{(c)}, \mathbf{p}^{(c)}, \alpha^{(c)}$.
   (b) Compute $\mathbf{d}^{(c+1)}$ by using on the data $(\mathbf{v}^1, \ldots, \mathbf{v}^r)$ the E-step starting from $\mathbf{c}^{(c)}, \mathbf{q}^{(c)}, \alpha^{(c)}$.
3. Compute $\boldsymbol{\theta}^{(c+1)} = (\mathbf{p}^{(c+1)}, \mathbf{q}^{(c+1)}, \alpha^{(c+1)})$
4. Repeat the steps (2) and (3) until convergence.

After we fit the mixture model to estimate $\boldsymbol{\theta}$, we can give an outright or hard clustering of this data by assigning each observation to the component of the mixture to which it has the highest posterior of probability of belonging. As the calculus of posterior probabilities starting form the parameter is not tractable, a simple solution is to use the probabilities c_{ik} and $d_{j\ell}$ obtained at the end of the block EM algorithm. This procedure, which assigns a partition to a value of the parameter $\boldsymbol{\theta}$, will be named " C-step " in the following.

3.2 CML approach and block CEM algorithm

With the CML approach, the partition is added to the parameters to be estimated. In [Govaert and Nadif, 2003], we have proposed the block CEM algorithm that is a variant of block EM. In each of the phases 2(a) and 2(b), it is sufficient to add a C-step which converts the c_{ik}'s and $d_{j\ell}$'s to a discrete classification before performing the M-step by assigning each object and each variable to cluster which has the highest posterior probability of belonging.

3.3 2EM and 2CEM algorithms

Obviously, we can also use the both classical versions EM and CEM on I and J separately (noted 2EM and 2CEM) but unfortunately it is unaware of the correspondence between I and J. It will be seen later that this process is ineffective to detect homogeneous blocs. In addition, the use of two models on the both sets is not parsimonious. Indeed, our proposed block mixture model has fewer parameters than a standard "one-dimensional" clustering: for example, with $n = 1000$ and $r = 500$ and equal proportions of mixture components, if we need to cluster binary data matrix into 4 clusters of rows and 3 clusters of columns, this leads to estimate 12 parameters with Bernoulli block mixture model instead of $5000 = 4 \times 500 + 3 \times 1000$ parameters with two Bernoulli mixture models, i.e., applied on I and J separately.

4 Numerical experiments

In this section, to illustrate the behaviors of our algorithms (2EM, 2CEM, block EM(1), block EM(2), block CEM) and to compare them, we studied their performances for the Bernoulli block mixture model where

$$\varphi(x; \alpha_{k\ell}) = (\alpha_{k\ell})^x (1 - \alpha_{k\ell})^{1-x} \text{ with } \alpha_{k\ell} \in]0, 1[.$$

With block EM(1) and block EM(2), we have two levels of convergence. The first is local; see the phases 2a) and 2b) for block EM(1) and the phase 2) for block EM(2) and the second convergence is global; see the phase (3) for block EM(1) and the phase (4) for block EM(2). In order to accelerate both algorithms, we decided to carry out less iterations locally and more at the global level. After intensive simulations, we chose to carry out only one iteration locally and considered that the global convergence is reached when $|1 - L^{(c)}/L^{(c-1)}| < \varepsilon$ where $L^{(c)}$ denotes the observed log-likelihood at c-th iteration and ε represents a threshold value which chosen on a pragmatic ground, here we took $\varepsilon = 10^{-7}$. This strategy, kept in the following, is fast and gives better results that when one chooses to carry out less iterations globally ($\varepsilon = 10^{-6}$) and more locally (This comparison is not reported here).

In our experiments, we selected twelve kinds of data arising from 3×2-component mixture model corresponding to three degrees of overlap (well

separated (+), moderately separated (++) or ill-separated (+++)) of the clusters and four sizes of the data ($n \times r = 50 \times 30, 100 \times 60, 200 \times 120, 300 \times 180$). The concept of cluster separation is difficult to visualize for Bernoulli-mixture models, but the degree of overlap can be measured by the Bayes error corresponding to the block mixture model. As its computation is being theoretically difficult, we used Monte Carlo simulations and evaluated the error rate by comparing the partitions simulated and those we obtained by applying a C-step. But, this step is not direct as in classical situation of mixture model and, in these simulations, we used a modified version of the block Classification EM algorithm in which the parameter $\boldsymbol{\theta}$ is fixed to the true value $\boldsymbol{\theta}^*$. Parameters have been chosen to obtain error rates respectively in $[0.01, 0.05]$ for the well-separated, in $[0.12, 0.17]$ for the moderately and in $[0.20, 0.24]$ for the ill-separated situations. For each of these twelve data structures, we generated 30 samples and for each sample, we have run five algorithms 20 times starting from the same random situations and selected the best solution for each method. We compared 2EM, 2CEM, block CEM, block EM(1) and block EM(2) with $(g, m) = (3, 2)$.

Firstly, we focused on the comparison between block EM(1) and block EM(2). To summarize the behavior of these algorithms, we computed the mean error rate and the mean running time for each simulation. From our results of experiments (Table 1), incontestably the both versions of block EM almost always give the same results and their performance increases with the size of data and especially for block EM(1) (with 300×180 and the situation +++ the error rate is equal to 0.22 for block EM(1) versus 0.28 for block EM(2)). We can also note that block EM(1) is faster and therefore a regular update of $\boldsymbol{\theta}$ is more advantageous. For the continuation, we kept only block EM(1).

The comparisons between 2EM, 2CEM, block CEM and block EM(1) are summarized in Table 2. The first one displays the mean error rate for each situation and in Table 3, the mean running time. From these experiments, the main point arising are the following.

- The versions 2EM and 2CEM working on the two sets separately are suitably effective only when the clusters are well separated. This shows the risk of the use of such methods to obtain homogeneous blocks.
- The block CEM algorithm, even if it is faster and better than 2CEM and 2EM does not give encouraging results when the clusters are not well separated. Moreover, when the size of data increases, it has some difficulties to detect the pattern into 3×2 blocks.
- Not surprisingly, the versions 2CEM and 2EM are slower than block CEM and block EM(1).

In our comparisons we chose to use the percentage of misclassified like an approximation of the Bayes error. This choice is justified because the number of obtained clusters and simulated ones were the same ones. Furthermore, we have extended these comparisons to the cases where the numbers of clusters

Size	Degree of overlap	Error rates		Times	
		block EM(1)	block EM(2)	block EM(1)	block EM(2)
(50,30)	+	.02(.02)	.02(.02)	0.11(0.07)	0.33(0.15)
	++	.24(.08)	.23(.09)	0.53(0.36)	1.71(1.26)
	+++	.31(.14)	.31(.13)	0.48(0.32)	2.04(1.53)
(100,60)	+	.02(.02)	.02(.02)	0.23(0.16)	0.77(0.70)
	++	.14(.03)	.14(.03)	0.28(0.13)	0.93(0.24)
	+++	.28(.11)	.28(.10)	0.69(0.51)	2.13(0.95)
(200,120)	+	.02(.01)	.02(.01)	0.42(0.08)	1.26(0.17)
	++	.14(.02)	.14(.02)	1.03(0.36)	3.72(0.89)
	+++	.28(.09)	.28(.09)	2.54(1.56)	9.86(4.09)
(300,180)	+	.03(.01)	.03(.01)	0.98(0.15)	3.43(0.30)
	++	.15(.02)	.15(.02)	3.11(2.66)	10.38(2.98)
	+++	.22(.06)	.28(.06)	3.90(1.72)	14.77(4.41)

Table 1. Means and standard errors (in parentheses) of error rates and times recorded from the 20 same random situations by block EM(1) and EM(2).

are different from $(3, 2)$ and used the Rand index in comparing the agreement between the both partitions (simulated and obtained). Note that this measure is not restricted to comparing partitions with the same number of clusters. The results of experiments have confirmed the performance of block EM(1).

Size	Degree of overlap	Error rates			
		2CEM(1)	2EM(2)	block CEM	block EM(1)
(50,30)	+	.09(.09)	.04(.06)	.02(.02)	.02(.02)
	++	.38(.08)	.31(.11)	.29(.11)	.24(.08)
	+++	.51(.13)	.46(.13)	.35(.12)	.31(.14)
(100,60)	+	.08(.06)	.07(.04)	.03(.02)	.02(.02)
	++	.31(.08)	.24(.09)	.16(.08)	.14(.03)
	+++	.53(.07)	.49(.10)	.35(.11)	.28(.11)
(200,120)	+	.03(.02)	.02(.01)	.02(.01)	.02(.01)
	++	.41(.10)	.29(.09)	.16(.08)	.14(.02)
	+++	.61(.07)	.50(.08)	.46(.10)	.28(.09)
(300,180)	+	.06(.02)	.05(.01)	.03(.01)	.03(.01)
	++	.50(.06)	.31(.06)	.15(.03)	.15(.02)
	+++	.58(.07)	.39(.08)	.37(.09)	.22(.06)

Table 2. Comparison between 2CEM, 2EM, block CEM, block EM(1): means and standard errors (in parentheses) of error rates.

Size	Degree of overlap	Times			
		2CEM	2EM	block CEM	block EM(1)
(50,30)	+	2.29(2.61)	0.53(0.12)	0.03(0.01)	0.11(0.07)
	++	0.23(0.02)	0.87(0.12)	0.10(0.21)	0.53(0.36)
	+++	0.37(0.83)	0.91(0.12)	0.07(0.12)	0.48(0.32)
(100,60)	+	2.19(0.48)	5.29(1.38)	0.39(0.25)	0.23(0.16)
	++	1.60(0.45)	6.97(0.99)	0.15(0.24)	0.28(0.13)
	+++	1.16(0.09)	7.71(1.09)	0.07(0.03)	0.69(0.51)
(200,120)	+	10.21(1.08)	26.49(9.14)	0.08(0.05)	0.42(0.08)
	++	10.12(0.73)	73.03(8.40)	0.19(0.10)	1.03(0.36)
	+++	8.97(0.80)	89.79(12.12)	0.21(0.12)	2.54(1.56)
(300,180)	+	37.31(2.77)	111.64(30.26)	0.27(0.27)	0.98(0.15)
	++	33.76(2.21)	291.01(31.84)	0.13(0.09)	3.11(2.66)
	+++	35.90(6.78)	449.28(407.16)	0.23(0.17)	3.90(1.72)

Table 3. Comparison between 2CEM, 2EM, block CEM, block EM(1): means and standard errors (in parentheses) of times recorded from the 20 same random situations.

5 Conclusion

Setting the problem of block clustering under the ML and CML approaches, we have compared three block clustering algorithms (block EM(1), block EM(2), block CEM) and two classical methods applied separately on the sets of rows and columns (2EM and 2CEM). Even if the both versions of block EM do not maximize exactly the likelihood, as in the classical mixture model situation but only an approximation of the likelihood of the block mixture model, they give encouraging results on simulated binary data and are better than the other methods. Furthermore, we note, that the first version block EM(1) appears slightly better than EM(2) when the clusters are ill-separated and it is faster. It would be now necessary to apply this algorithm to real situations and to extend this approach to other types of data, such as continuous data by using Gaussian densities for example.

References

[Cheng and Church, 2000]Y. Cheng and G. Church. Biclustering of expression data. In *Proceedings of ISMB*, pages 93–103, 2000.

[Dempster *et al.*, 1977]A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, pages 1–38, 1977.

[Dhillon, 2001]I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD Conference, San Francisco, California, USA*, pages 269–274, 2001.

[Govaert and Nadif, 2003]G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, pages 463–473, 2003.

[Govaert and Nadif, 2005]G. Govaert and M. Nadif. An EM algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 643–647, 2005.

[Symons, 1981]M.J. Symons. Clustering criteria and multivariate normal mixture. *Biometrics*, pages 387–397, 1981.