



Latent Block Model for Contingency Table

G rard Govaert¹ and Mohamed Nadif²

¹ HEUDIASYC, UMR CNRS 6599
Universit  de Technologie de Compi gne
BP 20529, 60205 Compi gne Cedex, France
(e-mail: gerard.govaert@utc.fr)

² CRIP5, Universit  Ren  Descartes
45, rue des Saint-P res,
75260 Paris Cedex 06, France
(e-mail: mohamed.nadif@univ-paris5.fr)

Abstract. Although many clustering procedures 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. This kind of methods has practical importance in a wide variety of applications such as text and market basket data analysis. Typically, the data that arises in these applications is arranged as two-way contingency table. Using Poisson distributions, a latent block model for these data is proposed and, setting it under the maximum likelihood approach and the classification maximum likelihood approach, various algorithms are proposed. Their performances are evaluated and compared to a simple use of EM or CEM applied separately on the rows and columns of the contingency table.

Keywords: Block Poisson mixture model, Block clustering, Contingency table, EM algorithm, CEM algorithm.

1 Introduction

Cluster analysis is an important tool in a variety of scientific areas including pattern recognition, information retrieval, micro-arrays and data mining. Although many clustering procedures such as hierarchical clustering and k -means, aim to construct an optimal partition of objects or, sometimes, variables, there are other methods, known as block clustering methods or latent block models, which consider the two sets simultaneously and organize the data into homogeneous blocks. Here, we restrict to block clustering methods defined by a partition of objects and a partition of variables.

A wide variety of procedures have been proposed for finding patterns in data matrices. These procedures differ in the pattern they seek, the types of data they apply to, and the assumptions on which they rest. In particular we should mention the work of [Hartigan, 1975], [Bock, 1979], [Govaert, 1983], [Govaert, 1984], [Govaert, 1995], [Arabie and Hubert, 1990]

and [Duffy and Quiroz, 1991] who have proposed some algorithms dedicated to different kinds of matrices. In recent years block clustering has become an important challenge in data mining. In the text mining field, [Dhillon, 2001] has proposed a spectral block clustering method which makes use of the clear duality between rows (documents) and columns (words). In the analysis of micro-array data, where data are often presented as matrices of expression levels of genes under different conditions, block clustering of genes and conditions has overcome the problem of the choice of similarity on the two sets, which occurs in conventional clustering methods [Cheng and Church, 2000].

The mixture model is undoubtedly a very useful contribution to clustering and offers considerable flexibility. Its associated estimators of posterior probabilities give rise to a fuzzy or hard clustering using the maximum a posteriori principle. To take into account the block clustering situation, we have developed a general *latent block model* [Govaert and Nadif, 2003]. In [Nadif and Govaert, 2005], a Poisson latent block model for two-way contingency table was proposed and the problem of clustering have been studied using the classification maximum likelihood approach (CML) leading to a block CEM algorithm.

In this paper, using the maximum likelihood setting, a block EM algorithm using an approximation of the likelihood was proposed and compared to block CEM, two-way EM and two-way CEM, i.e. EM and CEM applied separately on the rows and the columns of the data matrix.

The paper is organized as follows. In Section 2, the necessary background mixture approach of clustering using ML and CML approaches are given. The Poisson block model is defined in Section 3. In sections 4 and 5, EM and CEM algorithms associated to this model are proposed. Section 6 is devoted to numerical Monte Carlo experiments, and a final section summarizes and indicates the algorithm to be recommended.

We now define the notation that is used consistently throughout this paper. The two-way contingency table will be denoted \mathbf{x} ; it is a $n \times d$ data matrix defined by $\mathbf{x} = \{(x_{ij}); i \in I, j \in J\}$, where I is a categorical variable with n categories and J a categorical variable with d categories. A partition \mathbf{z} into g clusters of the sample I will be represented by the classification matrix $(z_{ik}, i = 1, \dots, n, k = 1, \dots, g)$ where $z_{ik} = 1$ if i belongs to cluster k and 0 otherwise. A similar notation will be used for a partition \mathbf{w} into m clusters of the set J .

Moreover, to simplify the notation, the sums and the products relating to rows, columns, row clusters and column clusters will be subscripted respectively by the letters i, j, k and ℓ , without indicating the limits of variation which will be implicit. So, for example, the sum \sum_i stands for $\sum_{i=1}^n$, and $\sum_{i,j,k,\ell}$ stands for $\sum_{i=1}^n \sum_{j=1}^d \sum_{k=1}^g \sum_{\ell=1}^m$.

2 Mixture Model

Before tackling the problem of block clustering we describe briefly two traditional approaches and the algorithms used to find the optimal partition. Some of the most popular heuristic clustering methods can be viewed as approximate estimations of probability models. For instance, the sum-of-squares criterion optimized by the k -means algorithm corresponds to the classification maximum likelihood criterion associated to a Gaussian mixture. We shall first provide a brief description of the mixture model.

In model-based clustering it is assumed that the data are generated by a mixture of underlying probability distributions, where each component k of the mixture represents a cluster. Thus, the data matrix is assumed to be an i.i.d sample $\mathbf{x}=(\mathbf{x}_1, \dots, \mathbf{x}_n)$ where $\mathbf{x}_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ from a probability distribution with density

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_k \pi_k \varphi(\mathbf{x}_i; \boldsymbol{\alpha}_k),$$

where $\varphi(\cdot; \boldsymbol{\alpha}_k)$ is the density of an observation \mathbf{x}_i from the k -th component and the $\boldsymbol{\alpha}_k$'s are the corresponding class parameters. The parameter π_k is the probability that an object belongs to the k -th component, and g , which is assumed to be known, is the number of components in the mixture. The parameter of this model is the vector $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ containing the mixing proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)$ and the vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_g)$ of parameters of each component. The mixture density of the observed data \mathbf{x} can be expressed as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \prod_i \sum_k \pi_k \varphi(\mathbf{x}_i; \boldsymbol{\alpha}_k).$$

The problem of clustering can be studied in the mixture model context using two different approaches: the maximum likelihood approach (ML) and the classification maximum likelihood approach (CML).

- The first approach estimates the parameters of the mixture, and the partition on I is derived from these parameters using the maximum a posteriori principle (MAP). The maximum likelihood estimation of the parameters results in an optimization of the log-likelihood of the observed sample. This optimization can be achieved using the EM algorithm.
- In the second approach, the partition is added to the parameters to be estimated. The maximum likelihood estimation of these new parameters results in an optimization of the complete data log-likelihood. This optimization can be performed using the Classification EM (CEM) algorithm [Celeux and Govaert, 1992].

Each of the two approaches has its advantages and its drawbacks. Through extensive simulation studies, comparisons between these approaches have

been made for continuous data by [Celeux and Govaert, 1993] and for binary data by [Govaert and Nadif, 1996].

To tackle the block clustering problem, we can use EM and CEM on I and J separately (from now on denoted 2EM and 2CEM), but to do so is to ignore the correspondence between I and J . A latent block model, which we examine in the following section, may be used to take this correspondence into account.

3 Latent Block Model

3.1 General model

For the classical mixture model, the pdf of a mixture sample $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ can be also written

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

where \mathcal{Z} denotes the set of all possible assignments \mathbf{z} of I into g clusters,

$$p(\mathbf{z}; \boldsymbol{\theta}) = \prod_{i,k} \pi_k^{z_{ik}} \quad \text{and} \quad f(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta}) = \prod_{i,k} \varphi(\mathbf{x}_i; \boldsymbol{\alpha}_k)^{z_{ik}}.$$

In the context of the block clustering problem, this formulation can be extended to propose a latent block model defined by the pdf

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{u} \in U} p(\mathbf{u}; \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{u}; \boldsymbol{\theta})$$

where U denotes the set of all possible assignments of $I \times J$, and $\boldsymbol{\theta}$ is the parameter of this mixture model.

In restricting this model to a set of assignments of $I \times J$ defined by a product of assignments of I and J , assumed to be independent, we obtain the pdf

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

where \mathcal{Z} and \mathcal{W} denote the sets of all possible assignments \mathbf{z} of I and \mathbf{w} of J . Now, as in latent class analysis, the $n \times d$ random variables x_{ij} are assumed to be independent once \mathbf{z} and \mathbf{w} are fixed; we then have

$$f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}},$$

where $\varphi(\cdot, \alpha_{k\ell})$ is a pdf defined on the real set \mathbb{R} .

3.2 Latent Block Model for Contingency table

When the data are a contingency table, we will assume that for each block $k\ell$ the values x_{ij} are distributed according the Poisson distribution $\mathcal{P}(\mu_i\nu_j\alpha_{k\ell})$ where the Poisson parameter is split into μ_i and ν_j the effects of the row i and the column j and $\alpha_{k\ell}$ the effect of the block $k\ell$. The pdf of this model can be written

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{z}, \mathbf{w} \in Z \times W} \prod_i \pi_{z_i} \prod_j \rho_{w_j} \prod_{i,j} \varphi(x_{ij}; \mu_i, \nu_j, \alpha_{z_i w_j}), \quad (1)$$

where

$$\varphi(x_{ij}; \mu_i, \nu_j, \alpha_{k\ell}) = \frac{e^{-\mu_i\nu_j\alpha_{k\ell}} (\mu_i\nu_j\alpha_{k\ell})^{x_{ij}}}{x_{ij}!},$$

$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\alpha})$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)$ and $\boldsymbol{\rho} = (\rho_1, \dots, \rho_m)$ are the vectors of probabilities π_k and ρ_ℓ that a row and a column belong to the k th component and to the ℓ th component respectively, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)$. In this work, the effects of the row i and the column j will be assumed to be known or estimated from the marginal totals.

Using this block model is dramatically more parsimonious than using a classical mixture model on each set I and J : for example, with $n = 1000$ rows and $d = 500$ columns and equal class probabilities $\pi_k = 1/g$ and $\rho_\ell = 1/m$, if we need to cluster the data matrix into $g = 4$ clusters of rows and $m = 3$ clusters of columns, the Poisson latent block model will involve the estimation of 12 parameters ($\alpha_{k\ell}, k = \{1, \dots, 4\}, \ell = \{1, \dots, 3\}$), instead of $(4 \times 500 + 3 \times 1000)$ parameters with two mixture models applied on I and J separately.

To cluster simultaneously the two sets I and J with this latent block model, we propose the use of the ML and CML approaches defined previously, as in the case of the classical mixture model.

4 ML approach for the latent block model

To apply the ML approach to the latent block model (1), the first step is the maximum likelihood estimation of the parameters and to solve this problem, we propose to use the EM algorithm. For this model, the complete data are taken to be the vector $(\mathbf{x}, \mathbf{z}, \mathbf{w})$ where unobservable vectors \mathbf{z} and \mathbf{w} are the labels. The EM algorithm maximizes the log-likelihood $L_M(\boldsymbol{\theta})$ w. r. to $\boldsymbol{\theta}$ iteratively by maximizing the conditional expectation of the complete data log-likelihood $L_C(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta})$ w. r. to $\boldsymbol{\theta}$ given a previous current estimate $\boldsymbol{\theta}^{(c)}$ and the observed data \mathbf{x} .

Unfortunately, difficulties arise owing to the dependence structure among the variables x_{ij} of the model. To solve this problem and using the

[Neal and Hinton, 1998] interpretation of the EM algorithm, we have proposed ([Govaert and Nadif, 2007]) to replace the maximization of the log-likelihood by the maximization of the fuzzy criterion

$$G(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta}) = L_C(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta}) + H(\mathbf{s}) + H(\mathbf{t})$$

where $\mathbf{s} = (s_{ik})$ with $s_{ik} \geq 0$ and $\sum_k s_{ik} = 1 \forall i$ and where $\mathbf{t} = (t_{j\ell})$ with $t_{j\ell} \geq 0$ and $\sum_\ell t_{j\ell} = 1 \forall j$.

In our situation, the maximization of this fuzzy criterion G can be performed with an alternated optimization using the following maximizations:

1. Maximization of $G(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta})$ w.r. to \mathbf{s} for fixed $\boldsymbol{\theta}$ and \mathbf{t} : it leads to the maximisation of $L_C(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta}) + H(\mathbf{s})$ which can be written

$$\sum_{i,k} s_{ik} (a_{ik} - \log s_{ik})$$

where

$$a_{ik} = \log \pi_k + \sum_\ell u_{i\ell} \log \alpha_{k\ell} - \mu_i \sum_\ell \nu_\ell \alpha_{k\ell},$$

$$u_{i\ell} = \sum_j t_{j\ell} x_{ij} \quad \text{and} \quad \nu_\ell = \sum_j t_{j\ell} \nu_j.$$

It can be easily shown that

$$s_{ik} = \frac{e^{a_{ik}}}{\sum_{k'} e^{a_{ik'}}} \quad \forall i.$$

2. Maximization of $G(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta})$ w.r. to \mathbf{t} for fixed $\boldsymbol{\theta}$ and \mathbf{s} : in a similar way, we obtain

$$t_{j\ell} = \frac{e^{b_{j\ell}}}{\sum_{\ell'} e^{b_{j\ell'}}} \quad \forall j$$

where

$$b_{j\ell} = \log \rho_\ell + \sum_k v_{jk} \log \alpha_{k\ell} - \nu_j \sum_k \mu_k \alpha_{k\ell},$$

$$v_{jk} = \sum_i s_{ik} x_{ij} \quad \text{and} \quad \mu_k = \sum_i s_{ik} \mu_i.$$

3. Maximization of $G(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta})$ w.r. to $\boldsymbol{\theta}$ for fixed \mathbf{s} and \mathbf{t} : it leads to the maximisation of $L_C(\mathbf{s}, \mathbf{t}, \boldsymbol{\theta})$ which can be written

$$\sum_k s_{k.} \log \pi_k + \sum_\ell t_{. \ell} \log \rho_\ell + \sum_{k,\ell} (y_{k\ell} \log \alpha_{k\ell} - \mu_k \nu_\ell \alpha_{k\ell})$$

where

$$s_{k.} = \sum_i s_{ik}, \quad t_{. \ell} = \sum_j t_{j\ell} \quad \text{and} \quad y_{k\ell} = \sum_{i,j} s_{ik} t_{j\ell} x_{ij}.$$

We obtain

$$\pi_k = \frac{s_{k.}}{n}, \quad \rho_\ell = \frac{t_{. \ell}}{d}, \quad \text{and} \quad \alpha_{k\ell} = \frac{y_{k\ell}}{\mu_k \nu_\ell}.$$

For this algorithm, denoted BEM in the following, 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 which it has the highest posterior probability of belonging to.

5 CML approach for the latent block model

To apply the CML approach to the latent block model (1), the partitions \mathbf{z} and \mathbf{w} are added to the parameters to be estimated, and the objective is to maximize the complete data log-likelihood associated with the latent block model. Unlike the ML approach, this maximization does not require an approximation and can be performed with alternated optimization.

This algorithm [Govaert and Nadif, 2003], termed BCEM in the following, can be viewed as a variant of BEM: in the step 1 and 2 of BEM, it is sufficient to add a C-step which converts the s_{ik} 's or $t_{j\ell}$'s to discrete classifications by assigning each row or each column to the cluster they have the highest posterior probability of belonging to. If we denote $\mathbf{z} = (z_{ik})$ and $\mathbf{w} = (w_{j\ell})$ the two partition matrices associated to these classifications, the step 3 is obtained by the maximization of $G(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta})$ w.r. to $\boldsymbol{\theta}$ for fixed \mathbf{z} and \mathbf{w} : it leads to the maximisation of $L_C(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta})$ which can be written

$$\sum_k z_{k.} \log \pi_k + \sum_\ell w_{. \ell} \log \rho_\ell + \sum_{k,\ell} (y_{k\ell} \log \alpha_{k\ell} - \mu_k \nu_\ell \alpha_{k\ell})$$

where

$$z_{k.} = \sum_i z_{ik}, \quad w_{. \ell} = \sum_j w_{j\ell} \quad \text{and} \quad y_{k\ell} = \sum_{i,j} z_{ik} w_{j\ell} x_{ij}.$$

We obtain

$$\pi_k = \frac{z_{k.}}{n}, \quad \rho_\ell = \frac{w_{. \ell}}{d}, \quad \text{and} \quad \alpha_{k\ell} = \frac{y_{k\ell}}{\mu_k \nu_\ell}.$$

6 Numerical experiments

In this section, to illustrate the behaviors of our algorithms, we studied their performances on simulated data. Four algorithms have been compared: the BEM and BCEM algorithms described in previous sections, and two-way EM and CEM algorithms, i.e. EM and CEM applied separately on I and J , denoting in the following 2EM and 2CEM.

In our experiments we selected twelve types of data arising from 3×2 -component mixture model corresponding to three degrees of cluster overlap

(well separated, moderately separated and ill-separated), and four data dimensions ($n \times d = 50 \times 30$, $n \times d = 100 \times 60$, $n \times d = 200 \times 120$ and $n \times d = 300 \times 180$).

The concept of cluster separation is difficult to visualize for Poisson latent block models, but the degree of overlap can be measured by the true error rate, which is defined as the average misclassification probability $E(\delta((\mathbf{z}, \mathbf{w}), r_B(\mathbf{x})))$ where r_B is the optimal Bayes rule

$$r_B(\mathbf{x}) = \operatorname{argmax}_{\mathbf{z}, \mathbf{w}} P(\mathbf{z}, \mathbf{w} | \mathbf{x})$$

associated to the latent block model and $\delta((\mathbf{z}, \mathbf{w}), (\mathbf{z}', \mathbf{w}'))$ is the proportion of misclassified items. Its computation being theoretically difficult, we used Monte Carlo simulations and approximated this error rate by comparing the partitions simulated with those we obtained by applying a classification step. Parameters were selected so as to obtain error rates respectively in $[0.04, 0.06]$ for the well-separated, in $[0.14, 0.16]$ for the moderately and in $[0.23, 0.26]$ for the ill-separated situations.

For each of these twelve data structures we generated 30 samples, for each sample we ran five algorithms 20 times starting from the same random situations, and we then selected the best solution for each method.

The simulation results are summarized in figures 1, 2 and 3. The first one displays the mean euclidean distance between true parameters and estimated parameters for each situation, the second the mean error rates and the third the mean execution times. The main findings arising from these experiments are the following: the 2EM and 2CEM algorithms processing the two sets separately are sufficiently effective only when the clusters are well separated, which shows that using these methods is risky; the BEM algorithm, even though it is slower than BCEM, gives generally the better results, especially when the clusters are not well separated; and not surprisingly, 2EM and BEM are slower than 2CEM and BCEM.

7 Conclusion

Placing the problem of block clustering within the ML and CML approaches, we have compared two block clustering algorithms (BEM and BCEM) and two classical methods applied separately to sets of rows and columns (2EM and 2CEM). Although block EM algorithm does not maximize the likelihood, as in the classical mixture model situation, but only maximizes a fuzzy criterion, this method gives encouraging results using simulated contingency data, and is better than the other methods. It would now appear desirable to apply this algorithm to real situations and to extend this approach to other types of data including continuous data, using Gaussian densities for example.

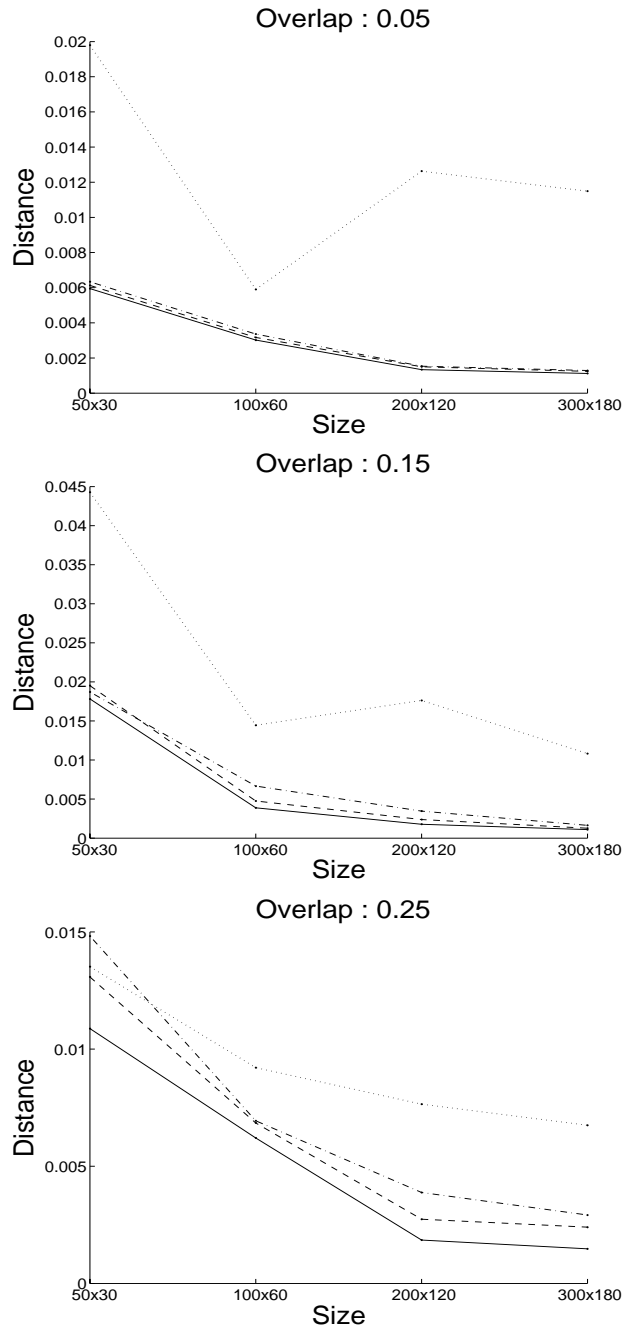


Fig. 1. Mean distance between true and estimated parameters for the 4 algorithms (BEM: solid line, BCEM: dashed line: 2CEM: dotted line and 2EM: dash-dot line) according to size and overlap.

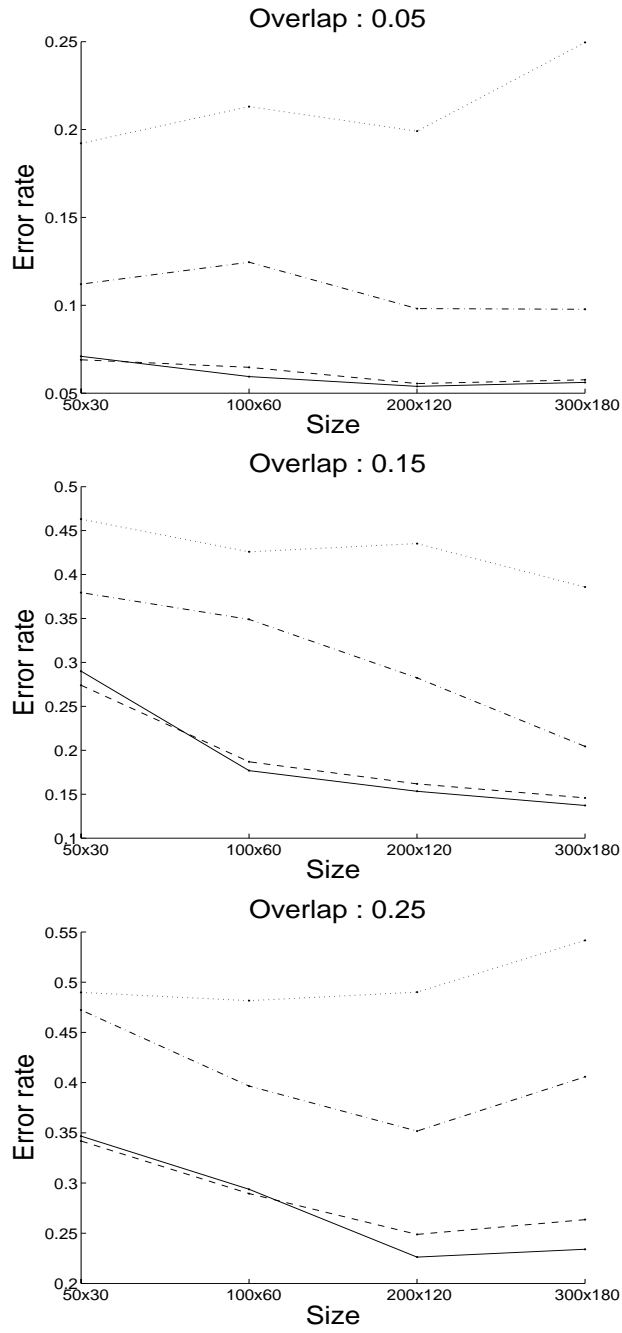


Fig. 2. Mean error rates for the 4 algorithms (BEM: solid line, BCEM: dashed line: 2CEM: dotted line and 2EM: dash-dot line) according to size and overlap.

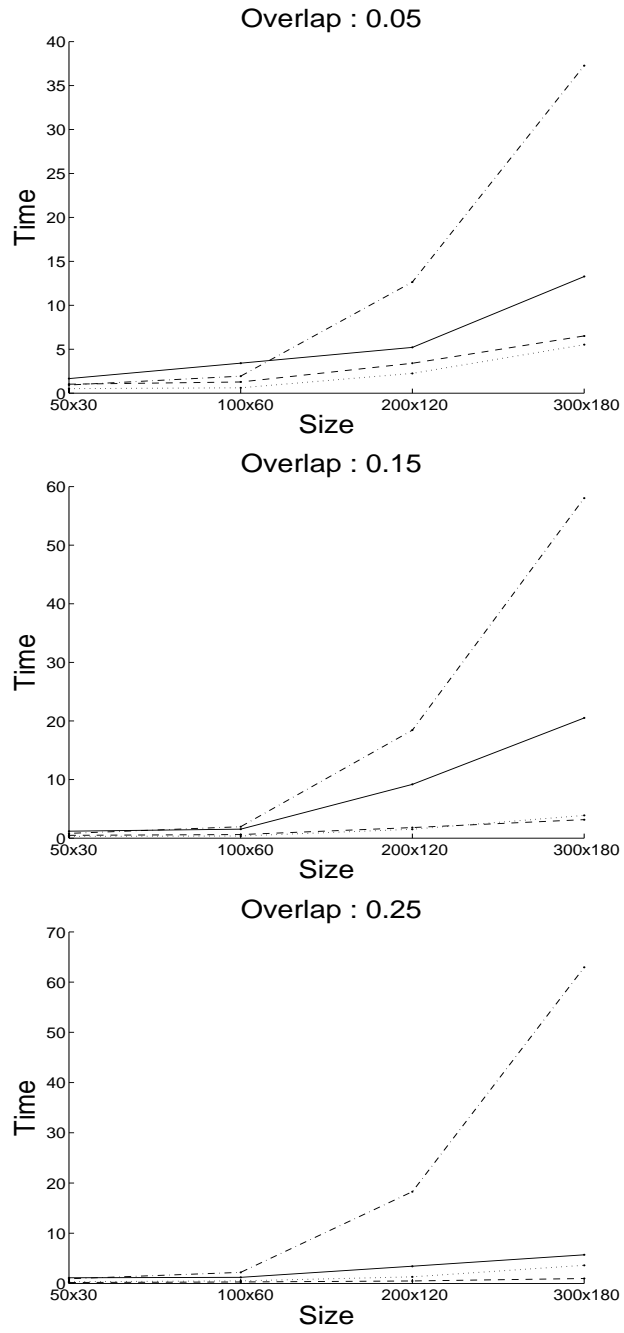


Fig. 3. Mean run time (in seconds) for the 4 algorithms (BEM: solid line, BEM: dashed line: 2CEM: dotted line and 2EM: dash-dot line) according to size and overlap.

References

- [Arabie and Hubert, 1990]P. Arabie and L. J. Hubert. The bond energy algorithm revisited. *IEEE Transactions on Systems, Man, and Cybernetics*, 20:268–274, 1990.
- [Bock, 1979]H. Bock. Simultaneous clustering of objects and variables. In E. Diday, editor, *Analyse des Données et Informatique*, pages 187–203. INRIA, 1979.
- [Celeux and Govaert, 1992]G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332, 1992.
- [Celeux and Govaert, 1993]G. Celeux and G. Govaert. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Statist. Comput. Simul.*, 47:127–146, 1993.
- [Cheng and Church, 2000]Y. Cheng and G. Church. Biclustering of expression data. In *ISMB2000, 8th International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, San Diego, California, August 19-23 2000.
- [Dhillon, 2001]I.S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Seventh ACM SIGKDD Conference*, pages 269–274, San Francisco, California, USA, 2001.
- [Duffy and Quiroz, 1991]D. E. Duffy and A. J. Quiroz. A permutation-based algorithm for block clustering. *Journal of Classification*, 8:65–91, 1991.
- [Govaert and Nadif, 1996]G. Govaert and M. Nadif. Comparison of the mixture and the classification maximum likelihood in cluster analysis when data are binary. *Computational Statistics and Data Analysis*, 23:65–81, 1996.
- [Govaert and Nadif, 2003]G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36:463–473, 2003.
- [Govaert and Nadif, 2007]G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, to appear, 2007.
- [Govaert, 1983]G. Govaert. *Classification croisée*. Thèse d’état, Université Paris 6, France, 1983.
- [Govaert, 1984]G. Govaert. Classification de tableaux binaires. In E. Diday, editor, *Data analysis and informatics 3*, pages 223–236, Amsterdam, 1984. North-Holland.
- [Govaert, 1995]G. Govaert. Simultaneous clustering of rows and columns. *Control and Cybernetics*, 24(4):437–458, 1995.
- [Hartigan, 1975]J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.
- [Nadif and Govaert, 2005]M. Nadif and G. Govaert. A comparison between block CEM and two-way CEM algorithm to cluster a contingency table. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, editors, *Knowledge Discovery in Databases, Lecture Notes on Artificial Intelligence (LNAI)*, number 3721, pages 609–616, Berlin Heidelberg, September 2005. Springer.
- [Neal and Hinton, 1998]Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–358. Kluwer Academic Publishers, Dordrecht, 1998.