ANR ClasSel

Modèles pour la classification croisée: Etat de l'art

Livrable 1.1

Contents

1 Introduction

In statistics and data analysis, the data often take the form of a rectangular table, that is an n by d data matrix $X = (x_{ij})$ defined on two sets I and J, sometimes referred to as two-way, two-mode data. For instance, I may be a set of cases (objects, persons), J may be a set of quantitative variables and the data matrix then collect the values taken by all the variables for each object. The sets I and J may be two categorical variables, the rows and columns then correspond to different categories of the two variables and the data matrix, which displays the frequency distribution of the variables, is the contingency table. The sets I and J may also be any two sets with a data matrix defining a binary relation on $I \times J$.

Given such a data matrix, the objective of data analysis can be viewed as the simultaneously analysis of the two sets I and J to identify underlying structures that may exist between these two sets. Different approaches such that exploratory analysis (graphical representation or numerical summary) or dimension reduction have been used. Principal component analysis and correspondence analysis are examples of such methods. This last method (Benzecri, 1973) is one of the best known methods that performs *simultaneously* analysis on both sets I and J. The table data must be a contingency table or at least have similar properties. The properties of this approach, especially transition formulas allow exchange the results of the tests on the sets I and J. These types of transitional help to define a set of relations of type barycentric justifying a simultaneous representation of two sets I and J. This representation allows to visualize simultaneously the proximity between the elements of I, the elements of J and the trends between I and J elements. Let us quote finally the methods of unfolding of Coombs (1950). The objective of these methods is to preferably represent a table on a line or a plan. The two sets are thus visualized simultaneously. Each individual is represented by an ideal point such as the relation of order between the variables defined by the distances in the ideal point in the various variables is closest to the order given in the table preferably initial.

Other methods relates to direct processing of the data matrix. For instance, seriation methods amounts to finding a permutation of rows associated with a permutation of columns leading to a reshaped data matrix with a maximum density of high cell values along the diagonal in addition to low value areas in the upper and lower parts. Such approaches have been used, for instance, in archaelogy, in phytosociology, in geography and in production management. Caraux (1984) proposed a criterion based upon an objective function with quadratic costs and Bertin (1980) proposed a manual heuristics based on visual densification. Factorial methods such as Benzecri's correspondence analysis (Benzecri, 1973) can also be used. When correspondence analysis gives rise to a U-shape effect ("Guttman effect") on the first two axes of the factorial representation, there exists a latent order

within the rows and the columns leading to diagonal band reshaping which corresponds to the order of the projections along the first axis of the rows and of the columns.

This survey is devoted to another family of methods leading a simultaneous analysis of two sets by using the notion of clustering. With a two-way two-mode data set, clustering algorithms are often applied to just one mode of the data matrix, which can be done in a hierarchical or non-hierarchical way. Among the non-hierarchical methods, k-means clustering (Hartigan, 1975a) is one of the most popular methods and has the advantage of a loss function being optimized. Contrary to this approach, there is a relatively new form of clustering trying to analyze the two sets simultaneously. These methods, named direct clustering, cross-clustering, simultaneous clustering, co-clustering, bi-clustering, two-way clustering, two-mode clustering or two-side clustering, has grown considerably in recent times.

A large number of such algorithms has been been proposed to date. One of the earliest and most cited biclustering formulation, known as block clustering, was proposed by Hartigan (1972, 1975b). He seeks to organize the data table using structures that may be, for example, defined from classifications on each of the two sets. This kind of methods are sometimes known as direct clustering. Older works may be cited. For instance, this problem was first described formally by Good (1965) which proposed a technique for simultaneous clustering of objects and variables. Fisher (1969) posed the problem of the simultaneous search for clustering on the row and column dimensions of a data matrix in a metric way. He defined a criterion to optimize, but offers no method to solve this problem. Tryon and Bailey (1970) first clusters variables using the correlation matrix and then, using a distance measure across the clusters of variables, clusters the cases. Dubin and Champoux (1970) proposed a method that combines the variables into types, and associates each individual to the types of variables forming a classification of individuals. More often, the authors discuss the classification of objects describing at length the choice of a measure of similarity and merely mention the possibility of a classification of variables without dealing on how to get there. Anderberg (1973) identified among the list of problems of classification the choice between I and J of all to classify. He considers as reasonable to classify variables as individuals. He even suggests an iterative approach in which the classification is done alternately on individuals and the variables until the classifications on individuals and the variables are mutually " harmonious"and believes that such research offers simultaneous " considerable potential to increase the effectiveness of automatic classification". In the case of contingency table and using as measure of information the χ^2 of contingency, Govaert (1977) developed the Croki2 algorithm to simultaneous search for partitions of each set minimizing the loss of information due to the regrouping in classes of the two sets. Extending this approach to binary, continuous and categorical data, he proposed (Govaert, 1983) the

Crobin, Croeuc and Cromul algorithms. Toledano and Brousse (1977) posed a similar problem: simultaneously build groups of individuals and groups of variables homogeneous between them and different ones compared to the others. Their objective is the search for two hierarchies checking this property. For this, they proposed an algorithm, called double aggregation which seeks with each iteration the best couple of lines or columns to be incorporated. Bock (1979) showed the interest of the simultaneous classification and gave several examples of problems for which a good solution is provided by a simultaneous classification.

Since that time this area has grown considerably and particularly in text mining and bioinformatics (Cheng and Church, 2000). An extensive overview of two-mode clustering methods can be found in (van Mechelen et al., 2004) and, in biological data analysis context, in (Madeira and Oliveira, 2004) and Prelic et al. (2006).

Section 2 is devoted to different types of data and applications that can be processed by biclustering. Section 3 summarizes the different approaches used in this area and Section 4 is dedicated to the model-based biclustering methods. The main software biclustering are listed in a section 5. A final section concludes this report.

2 Data and applications

2.1 Different types of data

In this survey, the data will always take the form of a rectangular table, that is an n by d data matrix $X = (x_{ij})$ defined on two sets I and J. This type of data, is sometimes referred to as two-way, two-mode data. Each element x_{ij} corresponds to a value representing the relation between row i and column j.

2.1.1 Object $\times$ variable data

In the most common situation, I corresponds to a set of n objects, each object being described by a set J of d variables. This type of data can be viewed as a sample of size n issued from a random variable of dimension d. Different type of variables can be used:

- Quantitative variables: the value taken by a quantitative variable are real numbers.
- Categorical variables: each of these variables comprises a set of discrete states or categories such that each object belongs to one and only one state.
- Binary variables: a particular case of categorical variable where there are only two states occurs frequently and merits separate consideration. Such two-state variables are considered as binary variables.

In this case, most authors distinguish two types of analysis: Tryon and Bailey (1970) speak of "0-Analysis" for the study of objects and "V-Analysis" for the study variables. According to them, the earliest works relate to the analysis of objects and this is the classification (taxonomy) and the first works on the analysis of the variables are from Pearson and Spearman and this is the factorial analysis. In other domains, these two types of analysis are called "P-technique" ant "Q-technique".

In this type of data, both sets show a strong asymmetry: the first corresponds to a sample consisting of n statistical units measured by d variables. There are other situations where the two sets play a similar role and can be interchanged. The most common example of this type of data corresponds to a contingency table.

2.1.2 Contingency table

There are many situations where one try to study the association between two categorical variables. In this case, the best way to do is to represent the data as a two-way contingency table (also referred to as cross-tabulation) which displays the frequency distribution of all the combinations of categories

of the two variables in a matrix format. Then, in this type of data matrix, the sets I and J are the categories of two categorical variables. A two-way contingency table is a way to summarize the two variables. We can remark that this definition can be easily extended to more categorical variables.

In this classical situation, the contingency table is computed starting from a sample of objects measured by the two categorical variables. There also exist situations where the observations are directly pairs $(i, j) \in I \times J$, where I and J are any sets (and not necessarily categorical variables) which lead to similar data, often called *dyadic data* or *co-ocurrence data* (COD). In this situation, the size of the sets I and J can be large and can lead to data sparseness.

This definition can also be extended to tables where every entry expresses a quantity of the same matter in such a way that all of the entries an be meaningfully summed up to a number to a number expressing the total amount of the matter in the data. Example of such data are trade tables showing the money transferred from i to j during a specified period.

2.1.3 Binary data

We find in many situations data matrices whose elements can take only two distinct values (yes/no, true/false, present/absent, agree/disagree,...). The data are then called *binary data*. This is what happens to object $\times$ variable data when all the variables are binary. Binary data can also be obtained as contingency tables where one retains only the presence or absence value. Such binary matrices, are found typically in ecology, whose one/zero entries respectively indicate possession/non-possession of a number of attributes (columns) by a sample of individuals (rows). In this situation, as for contingency data, the two sets I and J are treated symmetrically. In the following, the values of binary will be coded 0 or 1.

2.1.4 Continuous data

The sets of objects and variables are not comparable. We encounter the same problem with *principal component analysis* where the objects and the variables are not treated in a symmetrical way which is not the case of *correspondence analysis* which treats the rows and the columns of a co-occurrence matrix in the same way.

2.2 Data representations

Different representations can be associated to the data described in the previous section.

Geometrical representation For the object $\times$ quantitative variable data, a classical geometrical representation consists of regarding this data as

n points in d dimensions. In a dual way, the second representation, and less familiar, geometrical representation consists of regarding the data as d points in n dimensions. The classical methods like principal component analysis and k-means algorithm used extensively such representations. Similar geometrical representations can be used with contingency table. Correspondence analysis (Benzecri, 1973) are based on these representations.

Bipartite graph In all situations, it is possible to associate to the data matrix a *bipartite graph* whose vertices are the elements of $I \cup J$. For object×variables data and contingency data, the edges of the graph are the set of pairs $\{(i,j), i \in I, j \in J\}$ and they are weighted by corresponding entries x_{ij} in the data matrix. For binary data, the edges of the graph are the set of pairs $(i,j), i \in I, j \in J\, x_{ij} = 1$. In these situations, the adjacency matrix of the graph is the matrix $[0\mathbf{x}; \mathbf{x}'O]$

2.3 Applications

Text and Web mining In information retrieval systems, the model commonly used to represent the data is the bag-of-word or vector space model (Salton and McGill, 1983). A set of words is chosen from the set of all words in all documents. Each document is a vector in the feature space formed by this words. The vector entries can be frequencies of some other measures. Thus, the entire document collection may be represented by a word-by-document matrix whose rows correspond to words and column to documents. Generally, each document contains only a small number of words and hence, the data matrix is very sparse. Since the data dimension may be huge, a lower dimensional representation is imperative for efficient manipulation and biclustering is a reference tool to summarize the data.

Bioinformatics Gene expression data (Tibshirani et al., 1999) is defined by a large data matrix illustrating the expression levels of genes (rows of the matrix) under different samples such a tissues or experimental conditions (columns of the matrix). In this situation, the aim is to identify subsets of gene whose expression levels exhibit a coherent pattern under a subset of conditions.

Jagalur et al. (2007) use model-based block clustering to analyze a matrix of anatomy-by-gene expression level where each column correspond to the anatomical structures and rows correspond to genes.

Marketing The objective of recommender systems is to predict individual choices and preferences based on observed preference behavior. Collaborative filtering (Goldberg et al., 1992; Hofmann and Puzicha, 1999) is the method and process used to match data of one user with data for

similar users, based on purchase and browsing patterns. Collaborative filtering allows merchants to provide customers with future purchase recommendations. In this situation, biclustering can be a solution. For instance, for a recommending system in movie domain, because data are always sparse, much more accurate predictions can be made by grouping people into clusters with similar movies and grouping movies into clusters which tend to be liked by the same people.

Ecology Typically, in this domain the data take often the form of contingency data defined by the cover-abundance scores of a set of species in a set of sample units (quadrat, lake, county). Quite often, retaining only the information of presence or absence, the data take the form of a binary data (Podani and Feoli, 1991). It can be also a quantitative data. Bock (1979) cites an agricultural research institute that focuses on the performance of a set of varieties of fruits in different regions where these varieties are planted. The yields calculated for each variety and each region define a quantitative data. In all these situations, the biclustering allows to reduce the size of the data without losing too much information.

Group technology Group technology is a approach which has been widely used in many industries, including the design of job-shops and flexible manufacturing systems. Group technology is also very important for designing cellular manufacturing systems. In this situation, I is a set of n parts, J a set of d machines and x_{ij} is the processing time of part i using the jth machine. Cellular manufacturing involves processing a collection of similar parts (part families) on a dedicated cluster (or cell) of machines or manufacturing processes. This problem can be addressed by biclustering.

Archeology Leredde and Perin (1980) worked on a set of merovingian buckle-plates for which the presence or absence of a selection of criteria manufacturing techniques, shape and decoration has been observed. The problem was to structure the data by a series of permutations of rows and columns to show links between criteria and plates. The objective was to establish a typology of plates and criteria (biclustering problem) as well as highlight a temporal evolution in manufacturing techniques (seriation problem).

Computer science Schroeder (1977a,b, 1983) has used the Croki2 algorithm ((Govaert, 1977) to statistical approach to the study of program behavior via reference string analysis.

3 Different approaches

3.1 Definition of biclustering

Clustering can be defined as the process of organizing a set I of objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. More formally, the objective of clustering is to find a set $Z = \{Z_1, \ldots, Z_g\}$ where each Z_k, denoted cluster, is a subset of I. Different types of clustering are used. For example, Z can be a partition of I or a hierarchy or a set of overlapping clusters.

Extending this definition, *biclustering* can be defined as the process of organizing data matrix defined on two sets I and J into submatrices whose members are similar in some way. More formally, the objective of biclustering is to find a set $B = \{B_1, \ldots, B_b)\}$ where each $B_k = (Z_k, W_k)$, denoted bicluster or block, is the cartesian product of a subset Z_k of I and a subset W_k of J. Denoting $Z_1, \ldots, Z_g$ $(g \leq b)$ all the subsets Z_k with no repetition and in a similar way $W_1, \ldots, W_m$ $(m \leq b)$ all the subsets W_ℓ with no repetition, we obtain a clustering $Z_B = (Z_1, \ldots, Z_g)$ of I and clustering $W_B = (W_1, \ldots, W_m)$ of J. Different types of biclustering can be defined:

- B can be the Cartesian product of a partition Z of I and a partition W of J (two-mode partitioning); in this case, we have $Z = Z_B$ and $W = W_b$;
- B can be the Cartesian product of a hierarchy Z of I and a hierarchy W of J (two-mode hierarchical clustering);
- B can be a partition of $I \times J$;
- B can be a hierarchy of $I \times J$;

3.2 Two-mode partitionning

The simplest biclustering approach is to perform clustering of rows and clustering of columns using a partition Z of the set I of rows and a partition W of the set J of columns.

3.2.1 Introduction

While the goal is often the simultaneous study of two sets, many researchers have performed two-way clustering by applying algorithms to both sets separately and independently but with a simultaneous analysis of results. We can cite some examples of this type of approaches. In an article discussing the use of data analysis for architectural design Maroy and Peneau (1972) define their goal as "the study of the correspondences between object classes

and feature classes." For this, they perform a classification to obtain a partition of the objects and a partition of the characteristics. Then they examine the correspondence between the two classifications with the original table ordered according to an order respecting these partitions. We will return later to this concept of ordered array. The parallel use of correspondence analysis also allows them to study the links between the two classifications. Lerman and Leredde (1977) follow the same approach in an application on the characterization of file systems provided by different computer manufacturers. A partitioning method around nuclei (*pôle d'attraction*) is used to classify the two sets. The intersection of the two partitions obtained is then made to allow the reading and interpretation of results. In this study, again, the correspondence analysis is used in conjunction with the classification method. Tibshirani et al. (1999) illustrate several methods for two-way visualization of a reordered data matrix based on separately clustering genes and samples using two-way average linkage hierarchical clustering and two-way k-means clustering. We can fin similar approach for contingency data in Ciampi et al. (2005).

A more integrated approach is to classify one of the first sets and then, taking into account this classification, classifying the latter. Using the information bottleneck method introduced by Tishby et al. (1999) for finding the best tradeoff between accuracy and complexity (compression) when summarizing (e.g. clustering) a random variable X, Slonim et al. (2000) propose a two-stage clustering procedure for co-ocurrence data: the first stage uses a distribution clustering algorithm to obtain row clusters; in the second stage, these row clusters replace the original rows and a similar procedure is used to obtain column clusters.

But he most interesting situation consists of seeking both partitions simultaneously. As for the partitionning clustering situation, the most frequent approach consists to define a clustering criterion and then, to find an algorithm optimizing this criterion.

3.2.2 Clustering criteria

Z being a partition of the set I and W a partition of the set J, the problem is to find the couple (Z, W) optimizing $F(Z, W)$, F being a function which expresses the deviation existing between the couple (Z, W) and the initial table. The form of the criteria depends on the data.

Quantitative data When the data is a quantitative object×variable data, the most frequent criterion Govaert (1983, 1995) used is the least squares criterion which can be written

$$F(Z,W) = \sum_{k,\ell} \sum_{i \in Z_k, j \in W_\ell} (X_{ij} - \overline{x}_{k\ell})^2 = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (X_{ij} - \overline{x}_{k\ell})^2. \quad (1)$$

Here, $\overline{x}_{k\ell}$ is the mean of the submatrix defined by the clusters k and ℓ:

$$\overline{x}_{k\ell} = \frac{\sum_{i,j|z_{ik}=1,w_{j\ell}=1} X_{ij}}{z_k w_\ell}$$

where $z_k = \sum_i z_{ik}$ and $w_\ell = \sum_j w_{j\ell}$ are the cardinals of the clusters k and ℓ.

Adding a matrix $A = (a_{k\ell})$ where $a_{k\ell}$ is a value associated with each couple of classes k, ℓ, an extended version of this criterion can be defined in the following way

$$F(Z, W, A) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (X_{ij} - a_{k\ell})^2 \tag{2}$$

Two remarks can be made:

- For fixed partition Z and W, the optimal values $a_{k\ell}$ are the means $\overline{x}_{k\ell}$ and then, the optimal partitions for the two criteria are the same partitions;
- The matrix A, which has the same form as the initial data matrix X (real values), can be viewed as a summary of this matrix.

Bock (1979) extended this criterion and developed two variants: a no-interaction model

$$a_{k\ell} = \alpha + \beta_k + \gamma_\ell$$

and an interaction model

$$a_{k\ell} = \alpha + \beta_k + \gamma_\ell + \delta_{k\ell}.$$

In the first case, it is easy to show that the problem breaks up into two independent problems of search for partitions on each unit. The usual procedures of search for partitions can then be used. The optimal partition pair may be found by applying the one-way sum of squares clustering criterion separately to the rows and the columns. Thus the usual k-means procedure can be used. Introducing the values

$$Y_{ij} = X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..},$$

the second situation is equivalent to the criterion (1) with new values y_{ij}.

Contingency data For this type of data, the most common criteria are usually based on the concept of information measure such as the Pearson chi-square contingency or the mutual information.

If we note $s = \sum_{i,j} X_{ij}$ the sum of the contingency data, $f_{ij} = \frac{X_{ij}}{s}$ the relative frequencies, $f_{i.} = \sum_{j \in J} f_{ij}$ and $f_{.j} = \sum_{i \in I} f_{ij}$ the marginal frequencies, the chi-square can be written

$$\chi^2(I,J) = s \sum_{i \in I} \sum_{j \in J} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}.$$

This measure, represents the deviation between the theoretical frequencies $f_{i.}f_{.j}$, that we would have if I and J were independent, and the observed frequencies f_{ij}, usually provides statistical evidence of a significant association, or dependence between rows columns of the table. If I and J are independent the χ^2 will be zero and if there is a strong relationship between I and J, the χ^2 will be high. So, a significant chi-square indicates a departure from row or column homogeneity and can be used as a measure of the information brought by a contingency table. Various methods (Goodman, 1985) have been proposed for investigating this association. Some of them are graphical approaches and the best known is correspondence analysis.

Given a couple of partitions (Z, W), a new contingency data $A = (a_{k\ell})$ can be defined by regrouping the rows and columns according the partitions

$$a_{k\ell} = \sum_{i \in Z_k} \sum_{j \in W_\ell} X_{ij} \qquad \forall k = 1, \ldots, g \quad \text{and} \quad \forall \ell = 1, \ldots, m$$

and it can be shown that the chi-square $\chi^2(Z,W)$ associated to this new contingency table A verify

$$\chi^2(I,J) > \chi^2(Z,W).$$

Then, regrouping the elements of each cluster leads to a loss of the information and a natural objective will be to search for the partitions which minimize this lost. This leads to the maximization of the criterion

$$F(Z,W) = \chi^2(Z,W). \tag{3}$$

A similar development can be mase starting from the mutual information, also called Goodman RxC association,

$$G(I,J) = \sum_{i,j} f_{ij} \ln \frac{f_{ij}}{f_{i.}f_{.j}}.$$

The two criteria $\chi^2(Z,W)$ and $G(Z,W)$ are very closed and gives similar results.

Binary data In this situation, a natural choice is to search for homogeneous bloc $B_{k\ell}$, i.e. blocks with a majority of one or a majority of 0. If we note $A = (a_{k\ell})$ the binary binary matrix of size (g, m) where $a_{k\ell}$ is the

modal value of the block $B_{k\ell}$, the objective will be to minimize the following criterion

$$F(Z,W) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |X_{ij} - a_{k\ell}|. \tag{4}$$

This is a direct extension of the well-known maximal predictive classification criterion proposed by Gower (1974).

3.2.3 Algorithms

The optimization of the previous criteria is a a NP-hard problem and heuristic algorithms must be used. Different approaches have been proposed.

Alternated optimization algorithms are the most frequent approach: for instantance, for the criteria (1), (3) and (4), Govaert (1977, 1983, 1995) has developed tree algorithms Croeuc, croki2 and Crobin which alternates between row and column partitioning until the criterion reaches a local optimum. Bock (1979) and Dhillon et al. (2003) proposed respectively for quantitative data and contingency data similar algorithms.

Many other algorithms have been proposed: For instance, sequential algorithm (Podani and Feoli, 1991), genetic algorithm (Hansohm, 2000), simulated annealing (Bryan et al., 2005), tabu search (van Rosmalen et al., 2009),...

Puolamäki et al. (2008) and Tibshirani et al. (1999) compared these approaches with the use of classical clustering algorithms applied separately on the two sets.

3.3 Two-mode hierarchical clustering

The two-mode partitionning approach that we have seen can be extended to hierarchical clustering and, as for partitions, the process can be done separately or simultaneously.

3.3.1 Separate clustering

It was in a study on joint use of classification and analysis of correspondences that Jambu (1976) research links between the two hierarchical classifications obtained from the two sets I and J. In this example on data showing a time budget, I represents a set of population types and J is a set of types of activities undertaken by the population I. The simultaneous analysis is made on the two hierarchical models using the notion of contribution. In particular, the contributions of elements of I classes built on J and each element of J classes built on I are defined. Greenacre (1988) proposed the same approach and provide a simple graphical procedure which is useful in interpreting a significant chi-square statistic of a contingency table. In a similar way, Camiz and Denimal (1998) analyze a two-way contingency table

cross-classification based on two hierarchies obtained by classical approach (ward method) on each set I and J.

3.3.2 Simultaneous clustering

Toledano and Brousse (1977), Corsten and Denis (1990), Eckes and Orlik (1993) proposed two-way hierarchical clustering. Toledano and Brousse (1977) posed a similar problem: simultaneously build groups of individuals and groups of variables homogeneous between them and different ones compared to the others. Their objective is the search for two hierarchies checking this property. For this, they proposed an algorithm, called double aggregation which seeks with each iteration the best couple of lines or columns to be incorporated.

3.4 Direct clustering, block clustering

In the earliest and most cited biclustering formulation, known as direct or block clustering, Hartigan (1972) defines three types of biclustering which, using the notation defined in section 3.1, can be written

- Three tree structure: B, Z_B and W_B are hierarchies; I and J;
- Partitioned response: B are hierarchy and Z_B and W_B are partitions;
- Three partitions : B, Z_B and W_B are partitions (it is the two-mode partitionning).

Using a stepwise divisive method, Hartigan develops an algorithm minimizing the criterion (1). Tibshirani et al. (1999) have added a backward pruning and devised a permutation-based method for deciding on the optimal number of blocks. Duffy and Quiroz (1991) proposed another permutation-based algorithm for the same type of structures such that this approach can be extend to a wide variety of data, including matrices of categorical data. Eckes and Orlik (1993) have developed an hierarchical agglomerative algorithm to obtain a hierarchy of blocks.

3.5 Biclustering

In this more general situation, the problem is to identify a set of biclusters $B_k = (Z_k, W_k)$ such that each bicluster B_k, which is a submatrix, satisfies some specific characteristics of homogeneity with no other condition and, for instance, biclusters can overlap. Note that a submatrix is considered as a bicluster if it presents a particular pattern. There is no definition of what these patterns are. The choice of considering a submatrix as a bicluster, is subjective and depends on the context. However there are some basic patterns that can be used to identify a bicluster. They are called constant, additive

and multiplicative models. In constant models, all values in a bicluster are equal. In additive and multiplicative models, there is an additive and multiplicative factor between rows and columns respectively. Biclusters can also be identified by a mixture of these three models. Biclusters can overlap on the rows and/or columns, or present a tree or checkerboard structure. This diversity in the nature of biclusters accounts for the fact that no biclustering algorithm can identify all types of biclusters.

Several biclustering algorithms have been developed and applied to microarray analysis (Busygin et al., 2008). A good survey of biclustering methods for biological data analysis has been published by (Madeira and Oliveira, 2004), they enumerated more than 15 used in this context. Cheng and Church (2000) were the first to propose an algorithm for this task. They considered that biclusters follow an additive model and use the mean squared residue (MSR) to measure the coherence of the genes and conditions in a bicluster. The MSR takes this form

$$S(Z_k, W_k) = \frac{1}{z_k w_k} \sum_{i \in Z_k, j \in W_k} (X_{ij} - x_{i.} - x_{.j} + x_{..})^2$$

where

$$a_{i.} = \frac{1}{z_k} \sum_{i \in Z_k} X_{ij}, a_{.j} = \frac{1}{w_k} \sum_{j \in W_k} X_{ij}, a_{..} = \frac{1}{z_k w_k} \sum_{i \in Z_k, j \in W_k} X_{ij},$$

z_k and w_k are the size of clusters Z_k and W_k. This algorithm identifies biclusters one by one. A submatrix B_k is a δ-bicluster if $S(Z_k, W_k) \leq \delta$ for some $\delta > 0$. Applied to *yeast cell cycle data*, Cheng and Church (2000) identified several biologically relevant biclusters. However the setting of a threshold δ requires some prior knowledge which depends on the dataset (**?**).

Note that in such a situation, no criterion can be defined and this approach has lead to many heuristic: (Oyanagi et al., 2001) (algorithm ping-pong), (Ihmels et al., 2002, 2004; Ihmels and Bergmann, 2004), (Ben-Dor et al., 2003), (Bergmann et al., 2003), (Tanay et al., 2004), (Tchagang and Tewfik, 2006). Showing a connection between spectral partition and crossing minimization, Ahmad and Khokhar (2007) developed an efficient biclustering clustering.

Furthermore, Lazzeroni and Owen (2002) have proposed the popular plaid model. They assume that biclusters are organized in layers and follow a given statistical model incorporating additive two way ANOVA models. The search approach is iterative: Once $K-1$ layers (biclusters) have been identified, the K^{th} bicluster that minimizes a merit function depending on all layers is selected. They also applied their method to yeast data and found that genes in same biclusters share biological functions. Kluger et al. (2003) used a spectral approach for biclustering assuming that the data matrix contains a checkerboard structure after normalization. This structure is identified by a

singular value decomposition. They applied their method to *Lymphoma and Leukemia* datasets which contained different subtypes of cancer. On both datasets, conditions of the same subtype have been grouped together into the same biclusters. Tanay et al. (2002) have developed *SAMBA*, an approach based on the graph theory coupled with statistical modeling of the data. SAMBA, applied to a lymphoma dataset, produces biclusters representing new concrete biological associations. Cheng et al. (2008) have proposed the *pCluster* method that has the advantage to identify both additive and multiplicative biclusters in presence of overlap. They validated their method on yeast cell-cycle dataset using Gene Ontology annotations. Prelic et al. (2006) made a comparative study of different biclustering methods for gene expression. They used a very simple divide and conquer the *Bimax* algorithm as a reference to investigate the usefulness of different biclustering algorithms. They concluded that Bimax produces results similar to those of more complex methods.

The list of cited methods is not exhaustive, other approaches and methods were proposed. Abdullah and Hussain (2006) developed a graph-drawing-based biclustering technique based on the crossing minimization paradigm. Cano et al. (2007) proposed an extension to a possibilistic spectral algorithm, based on fuzzy and spectral clustering, allowing to obtain potentially overlapping biclusters. Finally, due to the increasing importance of the biclustering analysis of time series gene expression data, some algorithms such as CCC-Biclustering (Madeira and Oliveira, 2009; Madeira et al., 2010), have been proposed to address the problem of identification of biclusters with contiguous columns.

3.6 Others structures

3.6.1 Block diagonal structure

In the two-mode partitionning algorithms, constraints can be added to obtain a structure of diagonal blocks after row and column reorder: The row partition and the columns partition have the same number of clusters ($g = m$) and the diagonal biclusters (Z_k, W_k) take a form different from the other biclusters. For instance, in the binary data case, the diagonal biclusters will be composed primarily of value 1 and other biclusters of 0. The criterion (4) with the constraint that the matrix $A = (a_{k\ell})$ is the identity matrix is well adapted to this situation and has been used, for instance, by Garcia and Proth (1986) in a group technology application.

This type of approach is also known as *block seriation*. The techniques of seriation, met in various field such as sociology, archeology, botany, zoology, amount to finding a permutation of rows associated with a permutation of columns allowing to extract from the data a latent order. Block seriation corresponds to a particular approach of this problem and in this context,

Marcotorchino (1991) proposed a solution using linear programming. The Bond energy algorithm (BEA) (McCormick et al., 1972; Arabie and Hubert, 1990) can also be used to obtain a block diagonal structure.

Various other approaches have been proposed: Pensa et al. (2005) developed an algorithm leading to a block diagonal structure with the availability of overlapping. Dhillon (2001), modeling the document collection as a bipartite graph between documents and words, he considered the simultaneous clustering problem as a bipartite graph partitioning problem and obtained the clusters by using the second left and right singular vectors of the singular value decomposition of an appropriately scaled word-document matrix.

3.6.2 Different column clustering for each row cluster

Some authors (Pollard and van der Laan, 2002; Rocci and Vichi, 2008) have proposed to treat the two-mode clustering by first clustering the rows and then for each row cluster clustering the columns. For instance, in the partitionning situation, conditionnaly to each class of row partition, a different partition of columns is allowed.

3.6.3 Multi-way data

There exist more complex situations where the data take the form of a multidimensional array instead of a matrix. For example, multiple variables measured on a set of objects over time or contingency table defined on more than two categorical data are examples of array-valued or multiway data.

Some of the previous approaches have been extended to this situation. For instance, Ambroise and Govaert (2002) propose a clustering this multiway data along all its dimensions simultaneously using a model based strategy and Peng et al. (2008) propose the subspace clustering algorithm.

4 Co-clustering algorithms for Non-negative data matrices

Co-clustering methods become popular for dyadic data matrices such as co-occurrence matrix and binary data, arise frequently in market basket data or document clustering. The detection of homogeneous blocks in data matrix X can be reached by partitioning the rows into g clusters and the columns into m clusters. Different authors treated the co-clustering in a non-negative matrix factorization framework, others are considered to set the co-colustering in a mixture approach framework.

4.1 Non-negative matrix factorization

The non-negative matrix factorization (NMF) is useful for many applications in environment such as text mining, pattern recognition. NMF or two-factor factorization is one type of matrix factorizations. there are other types and the well know are semantic indexing (Berry et al., 1995), scaled PCA (Ding et al., 2002) and generalized SVD (Park and Howland, 2004). Now and frequently NMF is considered also as a clustering method and contrary to SVD, for example, it can still a latent semantic direction for each cluster.

Let be the non-negative arbitrary matrix X, in general NMF factorizes X into 2 arbitrary non-negative matrices $R \in \mathbb{R}_+^{n\times g}$ and $C \in \mathbb{R}_+^{\times d}$ and the goal is

$$\underset{R\geq 0, C\geq 0}{\text{Min}} ||X - RC^T||^2$$

It is proved by (Lee and Seung, 2001) that this error is non increasing under the iterative following updating rules

$$R_{ij} = R_{ij}\frac{(XC)_{ij}}{(RC^TC)_{ij}}$$

$$C_{ij} = C_{ij}\frac{(X^TR)_{ij}}{(CR^TR)_{ij}}.$$

The convergence of the iteration is guaranteed but the solution is not unique. If R and C constitute a solution, then for instance RD and CD^{-1} will also form another solution for any positive diagonal matrix D. To have the uniqueness solution, it suffices to require that Euclidean length of the column vector in R is one. This normalization leads to

$$R_{ij} = \frac{R_{ij}}{\sqrt{\sum_i R_{ij}^2}}$$

$$R_{ij} = C_{ij}\sqrt{\sum_i R_{ij}^2}.$$

Then each element R_{ij} represents the degree to which row belongs to cluster j, while each element C_{ij} of C indicates to which degree column is associated with cluster j. There is an equivalence between NMF (with I-divergence) (Lee and Seung, 2001) and probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999). The two methods optimize the same objective function. This fundamental fact and both and the L_1 normalization NMF ensure that NMF and PLSI are equivalent (Ding et al., 2008).

4.2 Non-negative Tri-factorization

Let be the non-negative arbitrary matrices $R = (R_{ik})_{n\times g}$, $C = (C_{j\ell})_{d\times m}$ and $A = (A_{k\ell})_{g\times m}$ designating respectively row and column memberships

and cluster representation which can be viewed as a summary of X. The problem is to look for these three matrices minimizing the total squared residue measure

$$F(R, C, A) = ||X - RAC^T||^2, \tag{5}$$

where $||.||$ denote Frobenius matrix norm and the superscript T denotes matrix transposition. The term RAC^T characterizes the information of X that can be described by the cluster structures. Then the clustering problem can be formulated as a matrix approximation problem where the clustering aim is to minimize the approximation error between the original data X and the reconstructed matrix based on the cluster structures. It can be formulated as a unconstrained 3-factor NMF

$$\underset{R\geq 0, A\geq 0, C\geq 0}{\text{Min}} ||X - RAC^T||^2$$

The approximation of X can be solved by an iterative alternating least-squares optimization procedures described below.

4.3 Non-negative block value decomposition: NBVD

As the objective function 5 is convex in R A and C respectively, but not convex in all of them simultaneously, it is not realistic to expect an algorithm to find the global minimum. The non-negative block value decomposition (NBVD) (Long et al., 2005) offers a solution of this problem by iteravely updating the decomposition using a set of multiplicative updating rules. This leads to have

$$R_{ij} = R_{ij} \frac{(XC^TA^T)_{ij}}{(RACC^TA^T)_{ij}}$$

$$A_{ij} = A_{ij} \frac{(R^TXC^T)_{ij}}{(R^TRACC^T)_{ij}}$$

$$C_{ij} = C_{ij} \frac{(A^TR^TX)_{ij}}{(A^TR^TRAC^T)_{ij}}$$

Furthermore, when A is identity matrix, this leads to the cluster model described by (Li, 2005) and (Xu et al., 2003). Note that the approximation of X by RAC^T is not unique and therefore does not offer directly a co-clustering of data. By assuming that RA is normalized to RAV, the cluster labels of the columns, are deduced by $V^{-1}C^T = (C_{ij})$; $w_{j\ell} = 1$ if $\ell = \text{argmax}_{\ell'=1,\ldots,m}\ C_{j\ell'}$ and $w_{j\ell} = 0$ otherwise. We can also deduce the label cluster rows by considering X^T.

4.4 Orthognal Non-negative Matrix tri-factorization

As noted before in classical NMF $X = RC^T$ there exist large number of matrices (A, B) such that $AB^T = I, RA \geq 0, CB \geq 0$; RA and BC is also the solution with the same residue. The orthogonality condition allows to overcome this difficulty, the formulation of the problem becomes

$$\underset{R\geq 0, C\geq 0}{\text{Min}} ||X - RC^T||^2, \text{ s.t } R^T R = I,$$

Note in this case, this optimization is equivalent to $k$means clustering. For co-clustering, it is therefore natural to consider to impose orthogonality on both R and C simultaneously in NMF.

$$\underset{R\geq 0, C\geq 0}{\text{Min}} ||X - RC^T||^2, \text{ s.t } R^T R = I, C^T C = I$$

This new formulation implies an equivalence with the simultaneous $k$means (Ding et al., 2005). However, this double orthogonality is very restrictive and gives poor approximation. Thus,one can consider

$$\underset{R\geq 0, A\geq 0, C\geq 0}{\text{Min}} ||X - RAC^T||^2, \text{ s.t } R^T R = I, C^T C = I$$

The unconstrained 3-factor NMF or tri-factorization is equivalent to unconstrained 2-factor ((Li and Zha, 2006),(Ding et al., 2006), (Yoo and Choi, 2010)). In their works, the authors studied the benefit of the orthogonality constraint to obtain rigorous clustering interpretation because the tri-factorization is interesting only when it cannot be transformed into 2-factor NMF. In **?**, the update rules in this case are the following

$$R_{ij} = R_{ij} \frac{(XCA^T)_{ij}}{(RR^T XCA^T)_{ij}}$$

$$A_{ij} = A_{ij} \frac{(R^T XC)_{ij}}{(R^T RAC^T C)_{ij}}$$

$$C_{ij} = C_{ij} \frac{(X^T RA)_{ij}}{(CC^T X^T RA)_{ij}}$$

Recent update rules exploiting true gradients on stiefel manifolds were also proposed by (Yoo and Choi, 2010)

$$R_{ij} = R_{ij} \frac{(XCA^T)_{ij}}{(RAC^T XR^T)_{ij}}$$

$$A_{ij} = A_{ij} \frac{(R^T XC)_{ij}}{(R^T RAC^T C)_{ij}}$$

$$C_{ij} = C_{ij} \frac{(X^T RA)_{ij}}{(CA^T R^T XC)_{ij}}$$

4.5 Co-clustering for binary data

By imposing some constraints on R, C and A, we can propose different criteria. For example, if R and C are two binary classification matrices noted $Z \in \{0,1\}^{n\times g}$ and $W \in \{0,1\}^{d\times m}$ and $A \in \{0,1\}^{g\times m}$, we can directly treat the co-clustering problem by minimizing

$$||X - ZAW^T||^2.$$

This criterion can be expressed as

$$\sum_{k,\ell} \sum_{i|z_{ik}=1} \sum_{j|w_{j\ell}=1} |X_{ij} - A_{k\ell}|.$$

and the problem of co-clustering can formulated as the following optimization problem:

$$\underset{Z,A,W}{\text{Min}} ||X - ZAW^T||^2, \text{ s.t } \sum_k z_{ik} = 1, \sum_\ell w_{j\ell} = 1$$

Different algorithms can be used to obtain a solution of this problem. Li (2005) has proposed an algorithm based on the use of the double $k$means principle. The principal steps are

1. Start from an initial position $(Z^{(0)}, W^{(0)}, A^{(0)})$.
2. Computation of $(Z^{(c+1)}, W^{(c+1)}, \mathbf{a}^{(c+1)})$ starting from $(Z^{(c)}, W^{(c)}, A^{(c)})$
   (a) Update $A^{(c+\frac{1}{2})}$: $A_{k\ell}^{(c+\frac{1}{2})} = \sum_{i,j} \frac{z_{ik}^{(c)} w_{j\ell}^{(c)} X_{ij}}{z_k^{(c)} w_\ell^{(c)}}$
   (b) Update $Z^{(c+1)}$, each i belongs to the kth cluster minimizing
   $$\sum_{j,\ell} w_{j\ell}^{(c)} (X_{ij} - A_{k\ell}^{(c+\frac{1}{2})})^2.$$
   (c) Update $W^{(c+1)}$, each j belongs to the ℓth cluster minimizing
   $$\sum_{i,k} z_{ik}^{(c)} (X_{ij} - A_{k\ell}^{(c+\frac{1}{2})})^2.$$
   (d) Computation of $A^{(c+1)}$ as in (a) step.
3. Iterate the steps 2 until the convergence.

Obviously the update of A can be performed before the update of W. This strategy appears more profitable because more faster. It was used by Govaert (1995) and moreover the author did not work on the original data

set but on intermediate matrices. In other words, in steps 2(b) and 2(c), for finding an optimal Z^{c+1} and W^{c+1}, the dynamic cluster algorithm proposed by (Diday and coll., 1980) is used to optimize the following criteria

$$F(Z, A|W) = \sum_{k} \sum_{i \in z_k} \sum_{\ell} |u_i^{\ell} - \#w_{\ell} A_{k\ell}|, \tag{6}$$

where $u_i^{\ell} = \sum_{j \in w_{\ell}} x_i^j$, and

$$F(W, A|Z) = \sum_{\ell} \sum_{j \in w_{\ell}} \sum_{k} |V_{kj} - \#z_k A_{k\ell}|, \tag{7}$$

where $V_{kj} = \sum_{i \in z_k} X_{ij}$ (# denotes the cardinality).

The step 2(b) is carried out by the application of the dynamic cluster algorithm using the $n \times m$ matrix $(U_{i\ell})$, the L_1 distance and kernels of the form $(\#w_1 A_{k1}, \ldots, \#w_m A_{km})$. Alternatively, the step 2(c) is carried out by the application of the dynamic cluster algorithm using the $g \times d$ matrix (V_{kj}), the L_1 distance and kernels of the form $(\#z_1 A_{1\ell}, \ldots, \#z_g A_{g\ell})$. Thus, at the convergence, we obtain homogeneous blocks of 0 or 1 by reorganizing rows and columns according to the partitions Z and W. Hence, each block $X_{k\ell}$, defined by the elements X_{ij} for $i \in z_k$ and $j \in w_{\ell}$ is characterized by A_k^{ℓ} which is the highest frequency value.

To help the user to interpret the results, some empirical statistics can be performed to evaluate the quality of the partition into blocks. For instance, we can define easily values $(1 - \varepsilon_{k\ell})$, each one of them corresponds to the proportion of block X_k^{ℓ} values equal to $A_{k\ell}$ and measures therefore the degree of homogeneity of $A_{k\ell}$.

One of the advantages of this kind of clustering methods is to summary the initial data matrix X in a simpler data matrix $(A_{k\ell})$ having the same structure. The data matrix $(A_{k\ell})$ is a binary matrix as the initial data matrix X. Moreover, this version is faster that the version proposed by (Li, 2005) and can process large data sets. Furthermore, the same algorithm with the χ^2 metric can be used to co-clustering the co-occurrence data Govaert (1995).

5 Model-based co-clustering

Clustering methods can be roughly divided into two categories. The first is based on a choice of some distance or distortion measure among the data points, which presumably reflects some background knowledge about the data. For most problems, a proper choice of the distance measure can be the main practical difficulty and through which much of the arbitrariness of the results can enter. Another class of methods is based on statistical assumptions on the origin of the data. Such assumptions enable the design of a statistical model where the model parameters are then estimated based on the given data and, in this situation, basing cluster analysis on mixture models has become a classical and powerful approach. (The work of Banfield and Raftery (1993), Celeux and Govaert (1992, 1993) and McLachlan (1982) are recent examples among many others of this point of view).

For the co-clustering, even if it is less common and more recent, various models have been proposed.

For instance, Rooth (1995) proposed a probabilistic model for block clustering of contingency data with a block diagonal structure and proposed an algorithm which uses formulas similar to the Baum-Welch re-estimation formulas for hidden Markov models. Hartigan (2000) used probabilistic models to perform block clustering on binary data. Nowicki and Snijders (2001) proposed a stochastic blockstructures model that builds a mixture model for stochastic relationships among objects and identifies the latent cluster via posterior inference. Kemp et al. (2006) proposed an infinite relational model that discovers stochastic structure in relational data in form of binary observations. Airoldi et al. (2008) proposed a mixed membership stochastic blockmodel that relaxes the single-latent-role restriction in stochastic block structures model. Govaert and Nadif (2003), proposed a latent block model defined by the following probability density function

$$f(X, \boldsymbol{\theta}) = \sum_{(Z,W)\in Z\times W} p(Z;\boldsymbol{\theta})p(W;\boldsymbol{\theta})f(X|Z,W;\boldsymbol{\theta}) \tag{8}$$

where Z and W denote the sets of all possibles assignments Z of objects and W of variables. In this model we also assume local independence i.e., the $n \times d$ random variables X_{ij} are assumed to be independent once Z and W are fixed; we have

$$f(X|Z,W;\boldsymbol{\theta}) = \prod_{i,j,k,\ell} \varphi(\mathbf{x}_{ij};\boldsymbol{\alpha}_k)^{z_{ik}w_{j\ell}}$$

where $\varphi(.;\alpha_{k\ell})$ is a probability density function defined on the real set $\mathbb{R}$. This model allows to propose algorithms for co-clustering binary and contingency tables by considering respectively Bernoulli and Poisson latent block models (see for instance; (Govaert and Nadif, 2008) and (Govaert and Nadif,

2010)). In each cases, variational approach, also named here mean-field approximation, of EM algorithm has been used to estimate the parameters and the partitions. Lashkari and Golland (2009) proposed the same generative model and also used variational approach. They showed that this model have common modeling assumptions with the Bregman coclustering of Banerjee et al. (2007). In analyzing continuous data in gene expression context, Jagalur et al. (2007) used the latent block model where the conditional distributions knowing the row and the column clusters are Gaussian. They applied the variational EM and the CEM algorithms. They also proposed a sequential optimization algorithm of the criterion defined in the CEM approach of the latent block model. For text categorization, Takamura and Matsumoto (2002) proposed a greedy algorithm to estimate simultaneously the parameters and the two partitions of the latent block model (classification approach) and used the Akaike criterion as stopping rule. In the contingency data situation, Hofmann et al. (1999) presented different clustering models and in the two-sided clustering situation, the modeling assumptions is equivalent to the relations obtained by the Poisson latent block model. They proposed an approximate EM algorithm using the variational approach and in the classification approach, they used a mutual information criterion.

To predict customer-product preference in market application, Deodhar and Ghosh (2007) proposed a model-based coclustering model which can be viewed as an extension of the latent block model taking into account attributes on customers and attributes on products. The proposed algorithm interleaves clustering of customers and products and construction of prediction models.

Different Bayesian approaches of this kind of models have been recently proposed. Shan and Banerjee (2008) and Dijk et al. (2009) developed a Bayesian approach of the latent block model to estimate the parameters. The first proposed a variational approach while the latter used Gibbs sampling algorithms. Always on the latent block model, similar works have been developed by Meeds and Roweis (2007) which showed how these models can easily take into account missing data and are robust to high rates of missing data. Starting from a probabilistic model on a contingency table, Poirier et al. (2008) also used a Bayesian approach to define a clustering criterion and proposed a greedy two-mode clustering algorithm to optimize this criterion.

In the collaborative filtering context, (Kleinberg and Sandler, 2008) proposed a mixture model and Ungar and Foster (1998) proposed a statistical model of collaborative filtering similar to the Bernoulli latent block model. For estimating the model parameters they have developed different methods including variations of k-means algorithm and Gibbs sampling.

Shafiei et al. (2006) extended these approaches and proposed a model-based overlapping coclustering able to work with any regular exponential family distribution.

Model-based approaches have also been used to treat the situation where different column clusterings are used for each row cluster For instance, in the analysis of gene expression data, Pollard and van der Laan (2002) proposed a probabilistic model and used the non parametric bootstrap method to assess the variability of the estimator ; in document and word clustering, Li and Zha (2006) used mixtures of Poisson distribution to model the multivariate distribution of the word counts in the document within each class.

6 Software

6.1 Blocks

The program BLOCKS is a program for stochastic block modeling, based on Snijders and Nowicki (1997) and Nowicki and Snijders (JASA, 2001). The method is based on Gibbs sampling, which is one of the many methods of Markov chain Monte Carlo. Therefore it is rather time-consuming.

This program can be used for undirected as well as directed graphs, but also for undirected or directed valued graphs (where you could think of 3 to 6 values).

The program is written in Delphi, for use under Windows (1995 and up). The current version is 1.7 (September 2006). BLOCKS is most easily executed from the StOCNET environment.

- Marc Flandreau and Clemens Jobst, "The Ties that Divide: A Network Analysis of the International Monetary System, 1890-1910", The Journal of Economic History, 65 (2005), 977-1007.
- Emmanuel Lazega, Saraï Sapulete, and Lise Mounier, Structural stability regardless of membership turnover? The added value of block modelling in the analysis of network evolution. Quality and Quantity, in press.

6.2 Bicat

BicAT (Bclustering Analysis Toolbox) (Barkow et al., 2006) is a freely available software written in Java which implements various biclustering methods as the algorithm of Cheng and Church (2000), the order-preserving submatrice algorithm of Ben-Dor et al. (2003) and the xmotifs algorithm of Murali and Kasif (2003). Functionalities as data handling, data preprocessing, data vizualization and postprocessng complement the software.

6.3 Biclust

Biclust (Kaiser and Leisch, 2008) is a R package for biclustering. The main function biclust provides several algorithms to find biclusters in two-dimensional data: Cheng and Church (2000), spectral biclustering (Kluger et al., 2003), Plaid Model (Lazzeroni and Owen, 2002; Turner et al., 2005), Xmotifs (Murali and Kasif, 2003) and Bimax (Prelic et al., 2006). In addition, the package provides methods for data preprocessing (normalization and discretisation), visualisation, and validation of bicluster solutions.

6.4 BiGGEsTS

BiGGEsTS (Madeira and Oliveira, 2009; Madeira et al., 2010) is a free open source software tool providing an integrated environnement for the biclustering of times series gene expression data. This software, coded in Java, enables a user friendly usage of the e-CCC-Biclustering algorithm and its extension in a graphical environnement together with the possibility to preprocess the data and postprocesss and anlyse the results using several criteria.

6.5 BiVisu

BiVisu (Cheng et al., 2007) is a software tool for bicluster detection and visualization.

6.6 Seriation

Seriation (Hahsler et al., 2009) is a R package for seriation.

6.7 Other software

Schepers and Hofmans (2009) propose a Matlab user interface (TwoMP) for two-mode partitioning on a given data by making use of an optimization algorithm supplemented by a validated model selection criterion.

7 Concluding remarks

7.1 Number of clusters

- see for instance the recent works of Schepers et al. (2008)
- Using the direct clustering algorithm of Hartigan (1972) which also produces hierarchical clustering trees for the rows and columns, Tibshirani et al. (1999) have added a backward pruning procedure and devised a permutation based method for deciding on the optimal number of blocks.

7.2 Initialization

- Spectral initialization procedure of Cho et al. (2004)

7.3 Others

- sparsity,
- empty clusters,
- model selection,
- problem complexity (large-high data),
- reorganization according to SOM and GTM

References

Abdullah, A. and Hussain, A. (2006). A new biclustering technique based on crossing minimization. *Neurocomputing*, 69:1882–1896.

Ahmad, W. and Khokhar, A. (2007). chawk: An efficient biclustering algorithm based on bipartite graph crossing minimization. In *VLDB 07*.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1823–1856.

Ambroise, C. and Govaert, G. (2002). A mixture model approach to datacube clustering (invited). In *26th Annual GFKL (Gesellschaft für Klassifikation)*, University of Mannheim, Germany.

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press, New York.

Arabie, P. and Hubert, L. J. (1990). The bond energy algorithm revisited. *IEEE Transactions on Systems, Man, and Cybernetics*, 20:268–274.

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., and Modha, D. S. (2007). A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *J. Mach. Learn. Res.*, 8:1919–1986.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821.

Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and Zitzler, E. (2006). Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283.

Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2003). Discovering local structure in gene expression data: The order-preserving submatrix problem. *Journal of Computational Biology*, 10(3-4):373–384.

Benzecri, J.-P. (1973). *L'analyse des données tome 2 : l'analyse des correspondances*. Dunod, Paris.

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physics Review E.*, 67.

Berry, M., Dumais, S., and W.O, G. (1995). Using linear algebra for intelligent information retrieval. *SIA review*, 37:573–595.

Bertin, J. (1980). Traitements graphiques et mathématiques. différence fondamentale et complémentarité. *Mathématiques et sciences humaines*, 72:60–71.

Bock, H. (1979). Simultaneous clustering of objects and variables. In Tomassone, R., editor, *Analyse des Données et Informatique*, pages 187–203, Le Chesnay, France,. INRIA.

Bryan, K., Cunningham, P., Bolshakova, N., Coll, T., and Dublin, I. (2005). Biclustering of expression data using simulated annealing. In *18th IEEE Symposium on Computer-Based Medical Systems, 2005. Proceedings*, pages 383–388.

Busygin, S., Prokopyev, O., and Pardalos, P. (2008). Biclustering in data mining. *Computers and Operations Research*, 35(9):2964–2987.

Camiz, S. and Denimal, J. (1998). A new method for cross-classification analysis of contingency data tables. In Payne, R. and P., G., editors, *Compstat 98 - Proceedings in Computational Statistics*, pages 209–214, Heidelberg. Physica-Verlag.

Cano, C., Adarve, L., and Blanco, A. (2007). Possibilistic approach for biclustering microarray data. *Computers in Biology and Medecine*, 37:1426–1436.

Caraux, G. (1984). Réorganisation et représentation visuelle d'une matrice de données numériques : Un algorithme itératif. *Revue de Statistique Appliquée*, 32(4).

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332.

Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Statist. Comput. Simul.*, 47:127–146.

Cheng, K., Law, N., Siu, W. C., and Lau, T. (2007). Bivisu: Software tool for bicluster detection and visualization. *Bioinformatics*, pages 1–2.

Cheng, K.-O., Law, N.-F., Siu, W.-C., and Liew, A. W. (2008). Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. *BMC Bioinformatics*, 9:210.

Cheng, Y. and Church, G. (2000). Biclustering of expression data. In *ISMB2000, 8th International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, San Diego, California. Disponible.

Cho, H., Dhillon, I., Guan, Y., and Sra, S. (2004). Minimum sum-squared residue co-clustering of gene expression data. In *Proceedings of The fourth SIAM International Conference on Data Mining*, pages 114–125.

Ciampi, A., Gonzalez Marcos, A., and Castejon Lima, M. (2005). Correspondence analysis and two-way clustering. *SORT, Statistics and Operations Research Transactions*, 29(1):27–42.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57:(3):145–158.

Corsten, L. and Denis, J.-B. (1990). Structuring interaction in two-way tables by clustering. *Biometrics*, 46(1):207–215.

Deodhar, M. and Ghosh, J. (2007). Simultaneous co-clustering and modeling of market data. In *Data Mining for Marketing Workshop, ICDM/07*.

Dhillon, I., Mallela, S., and Modha, D. (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pages 89–98.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, New York, NY, USA. ACM.

Diday, E. and coll. (1980). Optimisation en classification automatique. Le Chesnay INRIA.

Dijk, A. v., Rosmalen, J. v., and Paap, R. (2009). A bayesian approach to two-mode clustering. Econometric Institute Report EI 2009-06, Erasmus University Rotterdam, Econometric Institute.

Ding, C., He, X., and Simon, H. (2002). Unsupervised learning: self agregation in scaled principal component space. In *KDD'02*.

Ding, C., He, X., and Simon, H. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, pages 606–610.

Ding, C., Li, T., and Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52:3913–3927.

Ding, C., Li, T., Peng, W., and Park, H. (2006). Orthogonal nonnegative matrix tri-factorizations for clustering. In *In Proc SIGKDD Int'l Conf. on Knowledge Discovery and Data mining*.

Dubin, R. and Champoux, J. (1970). Typology of empirical attributes: Dissimilarity linkage analysis (dla). Technical report 3, University of California.

Duffy, D. E. and Quiroz, A. J. (1991). A permutation-based algorithm for block clustering. *Journal of Classification*, 8:65–91.

Eckes, T. and Orlik, P. (1993). An error variance approach to two-mode hierarchical clustering. *Journal of Classification*, 10(1):51–74.

Fisher, W. (1969). *Clustering and aggregation in economics.* Johns Hopkins Press.

Garcia, H. and Proth, J. M. (1986). A new cross-decomposition algorithm: The gpm comparison with the bond energy method. *Control and Cybernetics*, 15:155–165.

Goldberg, D., Nichols, D., Oki, B., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):70.

Good, I. J. (1965). Categorization of classification'in mathematics and computer science in biology and medicine. *HMSO, London*, pages 115–125.

Goodman, L. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, 13(1):10–69.

Govaert, G. (1977). Algorithme de classification d'un tableau de contingence. In *First international symposium on data analysis and informatics*, pages 487–500, Versailles. INRIA.

Govaert, G. (1983). *Classification croisée.* Thèse d'état, Université Paris 6, France.

Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics*, 24(4):437–458.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36:463–473.

Govaert, G. and Nadif, M. (2008). Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6):3233–3245.

Govaert, G. and Nadif, M. (2010). Latent block model for contingency tabl. *Communications in Statistics, Theory and Methods*, 3:416 425.

Gower, J. C. (1974). Maximal predictive classification. *Biometrics*, 30:643–654.

Greenacre, M. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, 75(3):457.

Hahsler, M., Hornik, K., and Buchta, C. (2009). Getting things in order: An introduction to the r package seriation. Technical report.

Hansohm, J. (2000). Two-mode clustering with genetic algorithms. In *24 th Annual Conference of the Gesellschaft fur Klassifikation*, pages 87–93.

Hartigan, J. (1975a). Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4(3):187–213.

Hartigan, J. (2000). Bloc voting in the united states senate. *Journal of Classification*, 17(1):29–49.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *JASA*, 67(337):123–129. Disponible.

Hartigan, J. A. (1975b). *Clustering Algorithms*. Wiley, New York.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *In: of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, number 289-296.

Hofmann, T. and Puzicha, J. (1999). Latent class models for collaborative filtering. In *IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 688–693, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hofmann, T., Puzicha, J., and Jordan, M. (1999). Unsupervised learning from dyadic data. *Advances in Neural Information Processing Systems*, 11.

Ihmels, J. and Bergmann, S. (2004). Challenges and prospects in the analysis of large-scale gene expression data. *Briefings in Bioinformatics*, 5(4):313–327.

Ihmels, J., Bergmann, S., and Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature genetics*, 31:370–377.

Jagalur, M., Pal, C., Learned-Miller, E., Zoeller, R. T., and Kulp, D. (2007). Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics*, 8:(Suppl 10):S5.

Jambu, M. (1976). Sur l'interprétation mutuelle d'une classification hiérarchique et d'une analyse des correspondances. *Revue de Statistique Appliquée*, 24(2):45–73.

Kaiser, S. and Leisch, F. (2008). A toolbox for bicluster analysis in R. In Brito, P., editor, *Compstat 2008—Proceedings in Computational Statistics*, pages 201–208. Physica Verlag, Heidelberg, Germany.

Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 381. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Kleinberg, J. and Sandler, M. (2008). Using mixture models for collaborative filtering. *Journal of Computer and System Sciences*, 74(1):49–69.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray cancer data: Co-clustering genes and conditions. *Genome Research*, 13:703–716.

Lashkari, D. and Golland, P. (2009). Co-clustering with generative models. Technical report, MIT.

Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86.

Lee, D. and Seung, H. (2001). Algorithms for non-negative matrix factorization. In *In Advances in Neural Information Processing Systems*, volume 13, pages 556–562.

Leredde, H. and Perin, P. (1980). Les plaques-boucles mérovingiennes. *Les dossiers de l'Archéologie*, 42:83–87.

Lerman, I. C. and Leredde (1977). La méthodes des pôles d'attraction. In *First international symposium on data analysis and informatics*, Versailles. North Holland.

Li, J. and Zha, H. (2006). Two-way poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics & Data Analysis*, 50(1):163 – 180. 2nd Special issue on Matrix Computations and Statistics.

Li, T. (2005). A general model for clustering binary data. In *KDD'05*, pages 188–197.

Long, B., Zhang, Z., and Yu, P. (2005). Co-clustering by value decomposition. In *KDD'05*, pages 635–640.

Madeira, S. and Oliveira, A. (2009). A polynomial time biclustering algorithm for finding genes with approximate expression patterns in gene expression time series. *Algorithms fo Molecular Biology*, 4(8).

Madeira, S., Teixeira, M., Sa-Correia, I., and Oliveira, A. (2010). Identification of regulatory modules in time series gene expression data using a linear time biclustering. *IEEE Transactions on Comput. Biology and Bioinformatics*, 7:153–165.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1):24–45. Disponible.

Marcotorchino, F. (1991). Seriation problems: an overview. *Applied Stochastic Models and Data Analysis*, 7(2):139–151.

Maroy, J.-P. and Peneau, J.-P. (1972). Analyse des données et conception en architecture. Bulletin 13, IRIA, Le Chesnay.

McCormick, W., Schweitzer, P., and White, T. (1972). Problem decomposition and data reorganization by a clustering technique. *Operations Research,*, 20(5):993–1009.

McLachlan, G. J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In Krishnaiah, P. R. and Kanal, L. N., editors, *Handbook of Statistics*, volume 2, pages 199–208, Amsterdam. North-Holland Publishing Company.

Meeds, E. and Roweis, S. (2007). Nonparametric bayesian biclustering. Technical Report UTML-TR-2007-001, University of Toronto.

Murali, T. M. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. In *Pac. Symp. Biocomput*, pages 77–88.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Oyanagi, S., Kubota, K., and Nakase, A. (2001). Application of matrix clustering to web log analysis and access prediction. In *WEBKDD 2001-Mining Web Log Data Across All Customers Touch Points, Third International Workshop*, pages 13–21.

Park, H. and Howland, P. (2004). Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*.

Peng, W., Li, T., and Shao, B. (2008). Clustering multi-way data via adaptive subspace iteration. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1519–1520, New York, NY, USA. ACM.

Pensa, R. G., Robardet, C., , and Boulicaut, J.-F. (2005). A bi-clustering framework for categorical data. In *Proceedings of the 9th European conferences on Principles and practice of Knowledge Discovery in Databases PKDD, LNCS*, volume 3721, pages 643–650. Springer-Verlag.

Podani, J. and Feoli, E. (1991). A general strategy for the simultaneous classification of variables and objects in ecological data tables. *Journal of Vegetation Science*, 2(4):435–444.

Poirier, D., Bothorel, C., and Boullé, M. (2008). Analyse exploratoire d'opinions cinématographiques: co-clustering de corpus textuels communautaires. In *Extraction et gestion des connaissances (EGC'2008)*, pages 565–576.

Pollard, K. S. and van der Laan, M. (2002). Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences*, 176(1):99–121. Disponible Recherche récusrvive de partitions sur chaque ensemble.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., and E., Z. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122 – 1129.

Puolamäki, K., Hanhijärvi, S., and Garriga, G. C. (2008). An approximation ratio for biclustering. *Information Processing Letters*, 108(2):45–49.

Rocci, R. and Vichi, M. (2008). Two-mode multi-partitioning. *Comput. Stat. Data Anal.*, 52(4):1984–2003.

Rooth, M. (1995). Two-dimensional clusters in grammatical relations. In AAAI Spring Symposium Series, S. U., editor, *Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*.

Salton, G. and McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill New York.

Schepers, J., Ceulemans, E., and Mechelen, I. (2008). Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. *Journal of Classification*, 25(1):67–85.

Schepers, J. and Hofmans, J. (2009). Twomp: A matlab graphical user interface for two-mode partitioning. *Behavior Research Methods*, 41(2):507–514.

Schroeder, A. (1977a). Étude de traces de références de programmes : analyse de processus à régimes. In *First International Symposium on Data Analysis and Informatics*, Rocquencourt. INRIA.

Schroeder, A. (1977b). A statistical approach to the study of program behaviour via reference strings analysis. In *Computer performances*, pages 381–396.

Schroeder, A. (1983). Une etude quantitative statique de programmes Pascal. 0 RT-0029, INRIA.

Shafiei, M., , and Milios, E. (2006). Model-based overlapping co-clustering. In *Proceedings of the Fourth Workshop on Text Mining, Sixth SIAM International Conference on Data Mining*, Bethesda, Maryland.

Shan, H. and Banerjee, A. (2008). Bayesian co-clustering. In *ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 530–539, Washington, DC, USA. IEEE Computer Society.

Slonim, N., Tishby, N., and Y, Y. I. (2000). Document clustering using word clusters via the information bottleneck method. In *In ACM SIGIR 2000*, pages 208–215. ACM press.

Takamura, H. and Matsumoto, Y. (2002). Two-dimensional clustering for text categorization. In *COLING-02: proceedings of the 6th conference on Natural language learning*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 Suppl 1:136–144.

Tanay, A., Sharan, R., and Shamir, R. (2004). Biclustering algorithms: A survey. *Handbook of Computational Molecular Biology*.

Tchagang, A. B. and Tewfik, A. H. (2006). Dna microarray data analysis: A novel biclustering algorithm approach. *EURASIP Journal on Applied Signal Processing*, 2006:1–12.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., and Brown, P. (1999). Clustering methods for the analysis of dna microarray data. Technical report, Department of Statistics, Stanford University.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.

Toledano, J. and Brousse, J. (1977). Une méthode de classification simultanée des lignes et des colonnes d'un tableau. In *1e Journées Internationales Analyse des Données et Informatique*, pages 105–107. IRIA.

Tryon, R. and Bailey, D. (1970). *Cluster Analysis*. McGraw-Hill, New York.

Turner, H., Bailey, T., and Krzanowski, W. (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational statistics & data analysis*, 48(2):235–254.

Ungar, L. and Foster, D. (1998). Clustering methods for collaborative filtering. In *AAAI Workshop on Recommendation Systems*, pages 112–125.

van Mechelen, I., Bock, H. H., and De Boeck, P. (2004). Two-mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, 13(5):363–394.

van Rosmalen, J., Groenen, P. J., Trejos, J., and Castillo, W. (2009). Optimization strategies for two-mode partitioning. *J. Classif.*, 26(2):155–181.

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *SIGIR'03*, pages 267–273.

Yoo, J. and Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information Processing and Management*, (559-570).