Co-Clustering

Co-Clustering

Models, Algorithms and Applications

Gérard Govaert
Mohamed Nadif

Series Editor
Francis Castanié

ISTE WILEY

First published 2014 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2013950131

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN: 978-1-84821-473-6

Printed and bound in Great Britain by CPI Group (UK) Ltd., Croydon, Surrey CR0 4YY

Table of Contents

Acknowledgment

This research was supported by the CLasSel ANR project ANR-08-EMER-002.

Introduction

Many of the data sets encountered in statistics are two dimensional and can be represented by a rectangular numeric table, that is an n by d data matrix $\mathbf{x} = (x_{ij})$ defined on two sets I and J, sometimes referred to as two-way or two-mode data. For instance, I may be a set of individuals (observations, cases, objects and persons) and J may be a set of variables (measurements, attributes and features). The data matrix then collects the values taken by all the variables for each individual. These data may be represented either as a table of individuals–variables as in the case of continuous variables, or as a frequency table or contingency table as in the case of categorical variables. In the following we examine a number of types of data on which co-clustering can be performed.

I.1. Types and representation of data

The type of a variable is determined by the set of possible values that the variable can take. In the following, we briefly review each type.

I.1.1. *Binary data*

Binary variables are widely used in statistics. Examples include presence–absence data in ecology, black and white pixels in image processing and the data obtained when recoding a table of qualitative variables. Data take the form of a sample $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ where $\boldsymbol{x}_i$ is a vector $(x_{i1}, \ldots, x_{id})$ of values x_{ij} belonging to the set $\{0, 1\}$. For example, the data might correspond to a set of 10 squares of woodland in which the presence (1) or absence (0) of two types of butterflies P1 and P2 was observed. Figure I.1 illustrates three alternative ways of presenting these data.

Square	P1	P2
1	1	1
2	1	0
3	1	1
4	0	1
5	0	1
6	1	1
7	1	0
8	0	0

State	Frequency
00	1
01	2
10	2
11	3

P1 \ P2	1	0
1	$n_{11} = 3$	$n_{10} = 2$
0	$n_{01} = 2$	$n_{00} = 1$

Figure I.1. *Example of binary data*

Binary data have been treated in clustering with a large number of distances, most of which are defined using the values n_{11}, n_{10}, n_{01} and n_{00} of the table crossing the two variables. For example, the distances between two binary vectors i and i' measured using "Jaccard's index" and the "agreement coefficient" can be written, respectively

$$\mathbf{d}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \frac{n_{11}}{n_{00} + n_{10} + n_{01}} \quad \text{and} \quad \mathbf{d}(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = n_{11} + n_{00}.$$

I.1.2. *Categorical data*

Categorical variables, sometimes known as qualitative variables or factors, are a generalization of binary data to situations where there are more than two possible values.

Here, each variable may take an arbitrary finite set of values, usually referred to as *categories*, *modalities* or *levels*. Like binary data, categorical data may be represented in different ways: as a table of individuals–variables of dimension (n, d), as a frequency vector for the different possible states, as a contingency table with d dimensions linking the categories or as a *complete disjunctive table* where categories are represented by their indicators. In this last form of representation, which we will use here, the data are composed of a sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, where $\boldsymbol{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$, with

$$\begin{cases} x_{ij}^h = 1 & \text{if } i \text{ takes the modality } h \text{ for the variable } j \\ x_{ij}^h = 0 & \text{otherwise,} \end{cases}$$

where m_j denotes the number of modalities of the variable j. In Figure I.2, a data matrix is shown which consists of a set of eight individuals described by three categorical variables A, B and C and its associated complete disjunctive table.

	A	B	C
1	1	1	1
2	2	1	1
3	1	3	1
4	2	2	1
5	2	2	2
6	1	3	2
7	1	1	3
8	1	1	3

	A1	A2	B1	B2	B3	C1	C2	C3
1	1	0	1	0	0	1	0	0
2	0	1	1	0	0	1	0	0
3	1	0	0	0	1	1	0	0
4	0	1	0	1	0	1	0	0
5	0	1	0	1	0	0	1	0
6	1	0	0	0	1	0	1	0
7	1	0	1	0	0	0	0	1
8	1	0	1	0	0	0	0	1

Figure I.2. *Example of categorical data (left) and its associated complete disjunctive table (right)*

I.1.3. *Continuous data*

Continuous data are undoubtedly the most current type of data and can be found in all areas. The structure takes the form of a relational table where the d columns are continuous variables and $\boldsymbol{x}_{ij} \in \mathbb{R}$. They can be positive or negative with

different units and variabilities. The measurement unit used can affect the results of different methods of data analysis and a normalization or transformation is often necessary. For instance, a variable can be normalized by scaling its values so that they lie within a specified range, such as $[0, 1]$. This aim can be achieved by the min-max normalization defined by

$$\frac{x_{ij} - \min_j}{\max_j - \min_j},$$

where $\min_j$ and $\max_j$ are, respectively, the lowest and the highest values taken by the variable j. The logarithmic transformation is also commonly used to pre-process data. These two transformations are frequently used with microarray data sets in order to overcome problems of inaccuracy of measurement or to provide values that are more easily interpretable. Other transformation techniques exist and are commonly used. We can cite, for instance, the z-score normalization defined by

$$\frac{x_{ij} - \mu_j}{\sigma_j},$$

where μ_j and σ_j are, respectively, the mean and the standard deviation of the variable j. Sometimes, and in order to reduce the effect of outliers, a variation of this z-score normalization consists of replacing σ_j by s_j, the mean absolute deviation of j. Different ways to normalize the data also exist. The user should pay special attention to this step as it is essential for obtaining meaningful results.

Besides, most authors distinguish two types of analysis: Tryon and Bailey [TRY 70] suggest “0-Analysis” for the study of objects and “V-Analysis” for the study of variables. According to them, the earliest works relate to the analysis of objects, which is the classification (taxonomy). The first work

on the analysis of the variables, from Pearson and Spearman, is the factor analysis. In other domains, these two types of analysis are called “P-technique” and “Q-technique”.

In the data previously described, both sets (individuals and variables) show a strong asymmetry, however in some situations the two sets play a similar role and can be interchanged. The contingency table studied in the next section is the most common example of this type of data.

I.1.4. *Contingency table*

There are many situations where we try to study the association between two categorical variables. A two-way contingency table is a method for summarizing the two variables. We can remark that this definition can be easily extended to more categorical variables. With data of this kind, the cells, formed by the cross-tabulation of two categorical variables, I having n categories and J having d categories, contain the frequency counts of the individuals belonging to these cells. Contingency tables of this sort can be found in many distinctive applications. An important example is information retrieval and document clustering, where I may correspond to a collection of documents and J to a set of words, the frequency denotes the number of occurrences of a word in a document. It is also noteworthy that the definition of the contingency table 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 can be meaningfully summed up to a number expressing the total amount of matter in the data. Examples of such data are trade tables showing the money transferred from country i to country j during a specified period. We now specify the notation that will be used to study the contingency table.

Let $\mathbf{x} = (x_{ij}, i = 1, \ldots, n; j = 1, \ldots, d)$ be a two-way contingency table associated with two categorical random variables that take values in sets $I = \{1, \ldots, n\}$ and $J = \{1, \ldots, d\}$. The entries x_{ij} are co-occurrences of row and column categories, each of which counts the number of entities that fall simultaneously into the corresponding row and column categories. The sum of frequencies of row and column categories, usually called marginals, are denoted by $x_{i.}$ and $x_{.j}$ and defined by $x_{i.} = \sum_j x_{ij}$, $x_{.j} = \sum_i x_{ij}$ and $x_{..} = \sum_{i,j} x_{ij}$. Here, we use the usual dot notation to express the sum with respect to the suffix replaced by a dot. Let $P_{IJ} = (p_{ij})$ denote the sample joint probability distribution. It is a matrix of size $n \times d$ defined by $p_{ij} = \frac{x_{ij}}{N}$ where $N = x_{..}$. The sample marginal probability distributions are defined by $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$. The sample joint probability distribution p_{ij} can be considered as estimators of the probabilities ξ_{ij} that the two categorical random variables occur in the cell in row i and column j. Table I.1 presents the form of the contingency table and of the corresponding sample joint distribution.

	1	$\ldots$	j	$\ldots$	d	
1	x_{1j}	$\ldots$	x_{1j}	$\ldots$	x_{1d}	$x_{1.}$
			$\vdots$	$\ldots$		
i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{id}	$x_{i.}$
			$\vdots$	$\ldots$		
n	x_{n1}	$\ldots$	x_{nj}	$\ldots$	x_{nd}	$x_{n.}$
	$x_{.1}$	$\ldots$	$x_{.j}$	$\ldots$	$x_{.d}$	N

	1	$\ldots$	j	$\ldots$	d	
1	p_{1j}	$\ldots$	p_{1j}	$\ldots$	p_{1d}	$p_{1.}$
			$\vdots$	$\ldots$		
i	p_{i1}	$\ldots$	p_{ij}	$\ldots$	p_{id}	$p_{i.}$
			$\vdots$	$\ldots$		
n	p_{n1}	$\ldots$	p_{nj}	$\ldots$	p_{nd}	$p_{n.}$
	$p_{.1}$	$\ldots$	$p_{.j}$	$\ldots$	$p_{.d}$	1

Table I.1. *Contingency table and sample joint distribution*

Sometimes, and specifically in document clustering when the rows are documents and the columns are words, some transformations of data are necessary. For instance, the co-occurrences can be replaced by the tf-idf statistics

[JON 72]. Different variants are proposed and commonly used in information retrieval and text mining.

I.1.5. *Data representations*

Different representations can be associated with the types of data described in the previous section.

Geometrical representation: for the continuous data, a classical geometrical representation consists of regarding these data as n points in d dimensions. In a dual way, a second and less familiar geometrical representation consists of regarding the data as d points in n dimensions. The classical methods, such as principal component analysis and k-means algorithm, used such representations extensively. Correspondence analysis [BEN 73b] uses similar geometrical representations to the contingency table.

Bipartite graph: in all situations, it is possible to associate the data matrix to a *bipartite graph* whose vertices are the elements of the union $I \cup J$ of sets I and J. For individuals$\times$variables table and the contingency table, the edges of the graph are the set of pairs $\{(i,j), i \in I, j \in J\}$ 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) such that $x_{ij} = 1$ (see, for instance, Figure I.3). This representation is frequently considered in the graph community such as in Web 2.0 tagging data and social networks.

The methods we are interested in next are clustering methods and, specifically, the simultaneous clustering of I and J. To this end, we will review the motivation of simultaneous analysis and then introduce co-clustering.

	B1	B2	B3	B4	B5	B6	B7
A1	1	1	1	0	0	1	1
A2	1	1	0	0	0	1	1
A3	0	0	0	1	1	1	1
A4	0	0	0	1	0	1	1

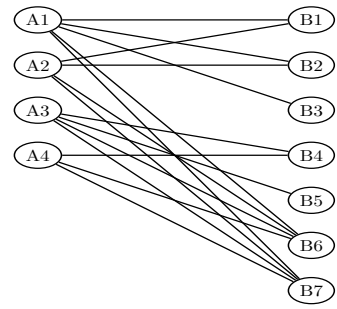

Figure I.3. *Binary data and its associated bipartite graph*

I.2. Simultaneous analysis

I.2.1. *Data analysis*

Given a data matrix, the objective of data analysis can be viewed as the simultaneous analysis of the two sets I and J to identify underlying structures that may exist between these two sets. Different approaches such as 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 given by Benzecri [BEN 73b] is one of the best known methods that performs analysis *simultaneously* on both sets I and J. The data table must be a contingency table or at least it must have similar properties. The properties of this approach, especially transition formulas, allow us to exchange the results on the sets I and J. These properties help us to define a set of barycentric relations, justifying a simultaneous representation of I and J and allowing us to simultaneously visualize the proximity among the elements of I, the elements of J and the elements of I and J. Finally let us quote the *unfolding method* of [COO 50] for which the objective is to represent rank preference data on a line or a plan. Each individual is represented by an ideal point such that the relation of order among the variables, defined by the distances between the

ideal point and the various variables, is closest to the order given in the initial data.

Other methods relate to direct processing of the data matrix. For instance, seriation methods amount 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 archaeology, phytosociology, geography and production management. Caraux [CAR 84] proposed a criterion based on an objective function with quadratic costs and Bertin [BER 80] proposed a manual heuristics based on visual densification. Factorial methods such as correspondence analysis can also be used. Note that when correspondence analysis gives rise to a U-shaped 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 columns.

This book is devoted to another group of methods of simultaneous analysis of two sets by using the notion of clustering. With a two-way or 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 [FOR 65, MAC 67, HAR 75b] is one of the most popular methods. Contrary to this approach, there is a relatively new form of clustering that analyzes the two sets simultaneously. These methods, called direct clustering, cross-clustering, simultaneous clustering, co-clustering, biclustering, two-way clustering, two-mode clustering or two-side clustering, have developed considerably in recent times.

I.2.2. *Co-clustering*

A large number of co-clustering algorithms have been proposed to date. One of the earliest and most cited biclustering formulations, known as block clustering, was proposed by Hartigan [HAR 72, HAR 75a]. He sought to organize the data table using structures that may be, for example, defined from classifications on each of the two sets. This kind of method is sometimes known as direct clustering. Older works can also be cited. For instance, this problem was first described formally by Good [GOO 65] who proposed a technique for the simultaneous clustering of objects and variables. Fisher [FIS 69] 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 for optimization, but offered no method to solve this problem. Tryon and Bailey [TRY 70] first clustered the set of variables using the correlation matrix and then, using a distance measure across the clusters of variables, clustered the set of individuals. Dubin and Champoux [DUB 70] proposed a method that combines the variables into types, and associates each individual with the types of variables forming a classification of individuals. More often, the authors discussed the classification of individuals, describing at length the choice of a measure of similarity and merely mentioned the possibility of a classification of variables without dwelling on how to get there. Anderberg [AND 73] identified the choice between I and J among the list of the problems of classification. He considered it reasonable to classify variables as individuals. He even suggested an iterative approach in which the classification is done alternately on the individuals and the variables until the classifications on both sets are mutually "harmonious", believing that such research simultaneously offers "considerable potential to increase the effectiveness of automatic classification".

In the case of contingency tables and using the chi-squared statistic χ^2 as measure of information, Govaert [GOV 77] developed an algorithm, called CROKI2 in the following to simultaneously search for partitions of each set, minimizing the loss of information due to regrouping the two sets into classes. Extending this approach to binary, continuous and categorical data, Govaert [GOV 83] proposed the CROBIN, CROEUC and CROMUL algorithms. Bock [BOC 79] 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 developed considerably and particularly in text mining (see, for instance, [DHI 03]). An extensive overview of two-mode clustering methods can be found in [VAN 04] and, in the context of biological data analysis, in [MAD 04].

I.2.3. *Applications*

Throughout the last four decades, biclustering has been used in many diverse areas. In this section, the aim is not to give an exhaustive list, but to briefly describe some applications and to provide some information about the situation of their use.

Text mining: in information retrieval systems, the model commonly used to represent the data is the bag-of-words or vector space model [SAL 83]. 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 these words. The vector entries can be frequencies or some other measures. Thus, the entire document collection may be represented by a word-by-document matrix whose rows correspond to words and columns 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 co-clustering is a reference tool to summarize the data. We can cite the works of co-clustering documents and words using bipartite spectral graph partitioning in [DHI 01] or information-theoretic in [DHI 03].

Web mining: web clustering is an approach for aggregating Web objects into various groups according to underlying relationships among them. The classical clustering algorithms are mainly used only on one dimension of the data, i.e. on user dimension or on page dimension, rather than taking into account the correlation among users and pages. Because the ultimate goal is to extract subsets of user sessions and Web pageviews to construct a variety of co-clusters, Web co-clustering is an effective means to address this challenge. Xu *et al.* [XU 10] proposed a procedure based on a bipartite spectral projection approach. To categorize the Web pages, Charrad *et al.* [CHA 09] used the CROKI2 co-clustering algorithm [GOV 77].

Bioinformatics: gene expression profiling has been established over the last two decades as a standard technique for obtaining a molecular fingerprint of tissues or cells in different biological conditions. The technology of microarrays enables us to measure mRNA levels for thousands of genes simultaneously [DER 96]. Given a set of gene expression profiles, organized as a large data matrix illustrating the expression levels of genes (rows of the matrix) under different samples, such as tissues or experimental conditions (columns of the matrix), a common aim is to identify subsets of gene whose expression levels exhibit a coherent pattern under a subset of conditions. Several works concern this topic and we will give a non-exhaustive list in section I.4.4. The interested reader can find a structured overview of the different algorithms used in [MAD 04] and [TAN 05].

Marketing: the objective of recommender systems is to predict individual choices and preferences based on observed preference behavior. Collaborative filtering is a technique used by some recommender systems, it consists of matching data of one user with data of similar users, based on purchasing and browsing patterns [GOL 92, HOF 99b]. Thus, it allows merchants to provide customers with future purchase recommendations. In this situation, biclustering is a good solution. For instance, for a recommender system in the movie domain, because data are always sparse, much more accurate predictions can be made by grouping people into clusters with similar movie preferences and grouping movies into clusters that tend to be liked by the same people. Current collaborative filtering techniques such as correlation and singular value decomposition methods are commonly used but are very expensive and can only be deployed in static off-line settings. To address these problems, in [GEO 05], the author proposes a collaborative filtering approach based on a weighted co-clustering algorithm [BAN 07] that involves the simultaneous clustering of users and items.

Ecology: in this domain the data often take the form of contingency tables defined by the cover-abundance scores of a set of species in a set of sample units (quadrat, lake and county). Quite often, retaining only the information of presence or absence, the data take the form of binary data [POD 91]. It can also be continuous data. Bock [BOC 79] cited an agricultural research institute that focused on the performance of a set of varieties of fruits planted in different regions. The yields calculated for each variety and each region define continuous data.

Group technology: group technology is an approach that was 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 is 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 co-clustering. In [GAR 86], the authors proposed a cross-decomposition algorithm called group production management (GPM) and showed that it outperforms the bond energy algorithm (BEA) introduced by McCormick *et al.* [MCC 72]. Note that in [MAR 87], the same objective was studied.

Archeology: Lerredde and Perin [LER 80] worked on a set of Merovingian buckle plates for which the presence or absence of a selection of criteria for manufacturing techniques, shape and decoration was 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 to highlight a temporal evolution in manufacturing techniques (seriation problem).

Computer science: to study the program behavior via reference string analysis, the CROKI2 co-clustering algorithm given by Govaert [GOV 77] was used in [SCH 77a], [SCH 77b] and [SCH 83]. Besides, the BEA was used for many applications such as imaging, ordering and related engineering problems and database attribute fragmentation [HOF 75, NAV 84]. In spatial databases, the cost of a spatial join can be very high due to the large size of spatial objects and the computation spatial operationally intensive. In [JIT 01], the authors proposed a variant of BEA to reduce the I/O cost of spatial-join processing. Recently, Bisson and Grimal [BIS 12] proposed a parallelizable approach for the co-clustering of multiview data sets.

I.3. Notation

In this section, we introduce notations that will be used throughout the book.

– 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, the sums $\sum_i$, $\sum_j$, $\sum_k$ and $\sum_\ell$ stand, respectively, for $\sum_{i=1}^{n}$, $\sum_{j=1}^{d}$, $\sum_{k=1}^{g}$, and $\sum_{\ell=1}^{m}$.

– In the same way, an $n \times d$ matrix $\mathbf{a}$ with elements a_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, d$ will simply be identified $\mathbf{a} = (a_{ij})$.

– Data will be denoted by an n by d data matrix $\mathbf{x} = (x_{ij}, i \in I = \{1, \ldots, n\}; j \in J = \{1, \ldots, d\})$.

– A partition of I into g clusters will be represented by the classification matrix $\mathbf{z} = (z_{ik}, i = 1, \ldots, n, k = 1, \ldots, g)$ where $z_{ik} = 1$ if element i arose from cluster k and $z_{ik} = 0$ otherwise. The kth cluster corresponds to the set of elements i such that $z_{ik} = 1$ and $z_{ik'} = 0 \; \forall k' \neq k$. Thus, the partition can be represented by a matrix of elements in $\{0, 1\}^g$ satisfying $\sum_k z_{ik} = 1$. For example, the partition made up of the two classes $\{1, 3, 4\}$ and $\{2, 5\}$ is denoted as

$$\mathbf{z} = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ z_{31} & z_{32} \\ z_{41} & z_{52} \\ z_{51} & z_{52} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Notation $(z_1, \ldots, z_n)$ where $z_i \in \{1, \ldots, g\}$ represents the cluster of i will also be used. Similarly, notation $\mathbf{w} = (w_{j\ell}, j = 1, \ldots, d, \ell = 1, \ldots, d)$ and $(w_1, \ldots, w_d)$ where $w_j \in \{1, \ldots, g\}$ represents the cluster of j will be used for partitions of J.

– In the same way, fuzzy partitions of I and J will be represented, respectively, by a matrix $\widetilde{\mathbf{z}}$ of elements in $[0, 1]$

satisfying $\sum_k \widetilde{z}_{ik} = 1$ for all $i = 1, \ldots, n$ and by matrix $\widetilde{\mathbf{w}}$ of elements in $[0, 1]$ satisfying $\sum_\ell \widetilde{\mathrm{w}}_{j\ell} = 1$ for all $j = 1, \ldots, d$. With the same example, we can have

$$\tilde{\mathbf{z}} = \begin{pmatrix} \widetilde{z}_{11} & \widetilde{z}_{12} \\ \widetilde{z}_{21} & \widetilde{z}_{22} \\ \widetilde{z}_{31} & \widetilde{z}_{32} \\ \widetilde{z}_{41} & \widetilde{z}_{52} \\ \widetilde{z}_{51} & \widetilde{z}_{52} \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \\ 0.7 & 0.3 \\ 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

– As for the contingency table, we use the usual dot notation to express the sum with respect to the suffix replaced by a dot. For instance, $x_{i.}$, $x_{.j}$ and $x_{..}$ will be used for data matrix $\mathbf{x}$, and $z_{.k}$ and $w_{.\ell}$ will represent the cardinalities of clusters.

– A^t denotes the transpose of matrix A.

– The use of the counting measure for discrete data allows us to unify the notation and "pdf" will be used in this book for both the probability density function in the continuous situation and the probability mass function in the discrete situation.

I.4. Different approaches

I.4.1. *Two-mode partitioning*

As we have seen in the previous section, a wide variety of procedures have been proposed for finding patterns in data matrices. These procedures differ in the patterns they seek, the types of data they apply to and the assumptions on which they are based. Here, we are concerned only with co-clustering defined by a partition of objects and a partition of variables.

I.4.1.1. *Example*

Clustering, which is the process of partitioning a set of objects or a set of variables, is a challenging research field.

All works on data mining devote a great part to clustering (see, for instance, [HAN 12a]). However, related problems with high-dimensionality and sparsity show the limits of clustering and in order to overcome them, two-mode partitioning offers a good solution. We can illustrate the significance of two-mode partitioning with a simple example. In Figure I.4(a), we present a binary data set described by the set of $n = 10$ objects I={A,B,C,D,E, F,G,H,I,J} and a set of $d = 7$ binary variables $J = \{1, 2, 3, 4, 5, 6, 7\}$. Figure I.4(b) consists of data reorganized by a partition of I into $g = 3$ clusters a={A,C,H}, b={B,F, J} and c={D,G, I,E}. Figure I.4(c) consists of data reorganized by the same partition of I and a partition of J into $m = 3$ clusters I={1, 4}, II ={3, 5, 7} and III ={2, 6}. Figure I.4(d) clearly reveals an interesting pattern and shows another advantage of co-clustering methods that reduce the initial data matrix x into a simpler data matrix with the same structure. In this example, our initial (10×7) binary data matrix is reduced to a $(g \times m) = (3 \times 3)$ summary binary data matrix. Note that different techniques using clustering algorithms on each set of data matrix can be used to reach the co-clustering objective; however, as we will see, these techniques reveal themselves to be less effective.

I.4.1.2. *Clustering toward two-mode clustering*

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 [MAR 72] defined their goal as "the study of the correspondences between object classes and feature classes". For this, they performed a classification to obtain a partition of the individuals and a partition of the characteristics. Then they examined the correspondence between the two classifications with the original table

arranged according to an order respecting these partitions. We will return to this concept of ordered arrays later. The parallel use of correspondence analysis also enabled them to study the links between the two classifications. Lerman and Leredde [LER 77] followed the same approach in an application on the characterization of file systems provided by different computer manufacturers. A partitioning method around the nuclei is used to classify the two sets. The intersection of the two partitions obtained is then made in order to allow the interpretation of results. In this study, again, the correspondence analysis is used in conjunction with the classification method. In bioinformatics, Tibshirani *et al*. [TIB 99] illustrated several methods for the 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 find a similar approach for contingency tables in [CIA 05].

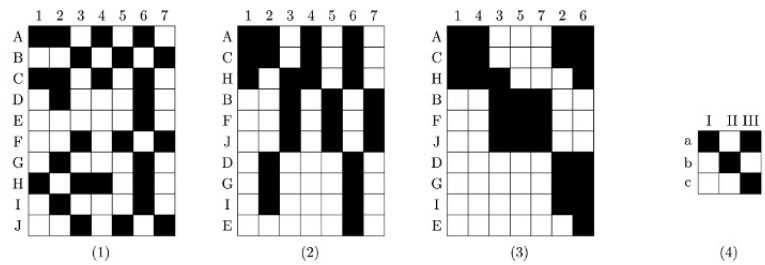

Figure I.4. *a) Initial binary data matrix. b) Data matrix reorganized according to a partition of rows. c) Data matrix reorganized according to partitions of rows and columns. d) Summary of this matrix*

As mentioned earlier, a more integrated approach is to classify one of the first sets and then, taking this classification into account, classifying the latter. Using the

information bottleneck method introduced by Tishby *et al.* [TIS 99] for finding the best trade-off between accuracy and complexity (compression) when summarizing (e.g. clustering) a random variable, Slonim and Tishby [SLO 00] proposed a two-stage clustering procedure for co-occurrence 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. However, the most interesting situation consists of seeking both partitions simultaneously.

I.4.1.3. *Criteria and algorithms*

As for the partitioning clustering situation, the most frequent approach consists of defining a clustering criterion and then finding an algorithm optimizing this criterion. Therefore, the problem is to find the couple of partitions $(\mathbf{z}, \mathbf{w})$, optimizing a function $F(\mathbf{z}, \mathbf{w})$, which expresses the deviation existing between the couple $(\mathbf{z}, \mathbf{w})$ and the initial data x. The form of the criteria depends on the data.

I.4.1.3.1. Continuous data

For continuous data, the most frequent criterion used is the least-squares criterion [GOV 83, GOV 95] that can be written as

$$F(\mathbf{z}, \mathbf{w}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - \overline{x}_{k\ell})^2, \qquad \text{[I.1]}$$

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

$$\overline{x}_{k\ell} = \frac{\sum_{i,j} z_{ik} w_{j\ell} x_{ij}}{z_{.k} w_{.\ell}}.$$

Adding a matrix $\mathbf{a} = (a_{k\ell})$ of size $(g \times m)$ 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

$$F(\mathbf{z}, \mathbf{w}, \mathbf{a}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - a_{k\ell})^2. \qquad [\text{I.2}]$$

Two remarks can be made: (1) for fixed partitions z and w, the optimal values $a_{k\ell}$ are the means $\overline{x}_{k\ell}$ and therefore, the optimal partitions for the two criteria are the same partitions. (2) 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. Furthermore, Bock [BOC 79] extended this criterion and developed two variants: a no-interaction model with $a_{k\ell}$ taking this form $a_{k\ell} = \alpha + \beta_k + \gamma_\ell$ and an interaction model with $a_{k\ell} = \alpha + \beta_k + \gamma_\ell + \delta_{k\ell}$. In the first case, it is easy to show that the problem divides into two independent problems of search for partitions on each unit. The usual procedures of search for partitions can therefore be used. The optimal partition pair may be found by applying the one-way sum of squares clustering criterion separately on 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 criterion [I.1] with new values y_{ij}.

Furthermore, there are a variety of methods for matrix approximation using matrix factorization as the principal component analysis for analyzing the rows and columns of a data matrix simultaneously. For example, we can cite the biclustering of gene expression data by non-smooth non-negative matrix factorization [CAR 06]. In fact, when x is

positive, even though the co-clustering problem is not the main objective of non-negative matrix factorization (NMF), this approach has attracted many authors for data co-clustering and particularly for document clustering. Given a non-negative matrix $\mathbf{x}$, different algorithms consist of seeking a three-factor decomposition $\mathbf{zaw}^t$ with all factor matrices restricted to be non-negative such as

$$\underset{\mathbf{z}\geq 0,\mathbf{a}\geq 0,\mathbf{w}\geq 0}{\arg\min} ||\mathbf{x}-\mathbf{zaw}^t||^2, \qquad \text{[I.3]}$$

where $||.||$ is the Frobenius norm.

The matrices $\mathbf{z}$ of size $(n \times g)$ and $\mathbf{w}$ of size $(d \times m)$ play the roles of row and column memberships and are not necessarily binary. Each value of both matrices $\mathbf{z}$ and $\mathbf{w}$ corresponds to the degree in which a row or column belongs to a cluster. The matrix $\mathbf{a}$ makes it possible to absorb the scales of $\mathbf{z}$, $\mathbf{w}$ and $\mathbf{x}$. All proposed algorithms are iterative, which can be differentiated by the update rules of the three matrices due to the chosen optimization method or the supplementary constraints imposed on the three matrices.

The approximation of $\mathbf{x}$ can be solved by an iterative alternating least-squares optimization procedure. For instance, the non-negative block value decomposition (NBVD) given by Long *et al.* [LON 05] offers a solution to this problem. Note that the solution to minimizing the criterion [I.3] is not unique. The authors proposed, at the convergence, to normalize each column of the matrix $\mathbf{za}$ to have the unit L^2 norm. The requirement of normalizing this matrix can be achieved using the new matrix $\mathbf{zab}$ where $\mathbf{b}$ is a diagonal matrix. The cluster labels of the columns are then deduced from the matrix $\mathbf{b}^{-1}\mathbf{w}^t$ instead of $\mathbf{w}^t$. We can also deduce the label cluster rows by working on the matrix $\mathbf{x}^t$. Note that in NBVD, only the non-negativity of the three matrices is required. In [DIN 06] and [YOO 10], the authors emphasized the importance of the orthogonality constraint, they introduced it on the matrices $\mathbf{z}$ and $\mathbf{w}$ and proposed, respectively, two variants of orthogonal non-negative matrix

tri-factorization called ONM3F in [DIN 06] and ONMTF in [YOO 10]. They can be differentiated by the update rules of the three factors. In a document clustering task, NBVD, ONM3F and ONMTF were shown to work well. Labiod and Nadif [LAB 11a] proposed a co-clustering framework based on NMF formulation. Contrary to previous approaches, they placed the co-clustering aim under the non-negative factorization at the beginning. The key idea is that the latent block structure in a rectangular non-negative data matrix is factorized into two factors rather than three factors, the row-coefficient matrix $\mathbf{r}$ and column-coefficient matrix $\mathbf{c}$ indicating, respectively, the degree in which a row and a column belong to a cluster. Then, under this framework they developed a variant co-clustering algorithm for non-negative data, which iteratively computes two factors based on two multiplicative update rules. The proposed approach optimizes a relaxed formulation of the double k-means criterion in an NMF style, then the optimization procedure looks for the best approximation of the matrix $\mathbf{x}$ by the matrix $\mathbf{r}\mathbf{r}^t\mathbf{x}\mathbf{c}\mathbf{c}^t$ with respect to some suitable constraints on the matrices $\mathbf{r}$ and $\mathbf{c}$.

I.4.1.3.2. Binary data

In this situation, a natural choice is to search for homogeneous blocks, i.e. blocks with a majority of ones or a majority of zeros. In this case, the objective is to minimize the criterion

$$F(\mathbf{z}, \mathbf{w}) = \sum_{i,j,k,\ell} z_{ik} w_{j\ell} |x_{ij} - a_{k\ell}|, \qquad \text{[I.4]}$$

where $a_{k\ell}$ is the modal value of the submatrix defined by the clusters k and ℓ. This criterion is a direct extension of the well-known maximal predictive classification criterion proposed by Gower [GOW 74].

Additionally, the non-negative matrix factorization approach has been used to identify biclustering in microarray

data. Zhang *et al.* [ZHA 10] extended standard NMF to binary matrix factorization (BMF) in order to solve the biclustering problem.

I.4.1.3.3. Contingency table

For this type of data, the most common criteria are usually based on the concept of information measures such as the phi-squared coefficient or mutual information (see Chapter 4). The phi-squared coefficient, which can be written as

$$\phi^2(I,J) = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}},$$

represents the deviation between the theoretical frequencies $p_{i.}p_{.j}$ that we would have whether I and J were independent and the observed frequencies p_{ij}. It usually provides statistical evidence of a significant association, or dependence among the rows and columns of the table. If I and J are independent, the ϕ^2 will be zero and if there is a strong relationship between I and J, the ϕ^2 will be high. So, a significant phi-squared coefficient 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 [GOO 85] have been proposed for investigating this association. Some of them are graphical approaches and the best known is the correspondence analysis.

Given a couple of partitions $(\mathbf{z}, \mathbf{w})$, a new contingency table $\mathbf{x}^{\mathbf{zw}} = (\mathbf{x}^{\mathbf{zw}}_{k\ell})$ can be defined by regrouping the rows and columns according to the partitions and it can be shown that the phi-squared coefficient $\phi^2(\mathbf{z}, \mathbf{w})$ associated with this new contingency table verifies

$$\phi^2(I,J) > \phi^2(\mathbf{z}, \mathbf{w}).$$

Therefore, regrouping the elements of each cluster leads to a loss of the information and a natural objective will be to

search for the partitions that minimize this loss. This leads to the maximization of the criterion

$$F(\mathbf{z}, \mathbf{w}) = \phi^2(\mathbf{z}, \mathbf{w}). \qquad [I.5]$$

A similar development can be made starting from the mutual information

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

As we will see in Chapter 4, the two criteria $\phi^2(\mathbf{z}, \mathbf{w})$ and $\mathcal{I}(\mathbf{z}, \mathbf{w})$ are very close and give similar results in most situations.

I.4.1.4. *Algorithms*

The optimization of the previous criteria is an NP-hard problem and heuristic algorithms must be used. Different approaches have been proposed. Alternated optimization algorithms are the most frequent approach: for instance, for criteria [I.2], [I.4] and [I.5], Govaert [GOV 77, GOV 83, GOV 95] developed three algorithms CROEUC, CROBIN and CROKI2, which alternate between row and column partitioning until the criterion reaches a local optimum. Bock [BOC 79] and Dhillon *et al.* [DHI 03] proposed similar algorithms, respectively, for continuous data and contingency table. These methods are fast and can process large data sets. Far less computation is required than for processing the two sets separately because, as we will see, the clustering is performed on reduced intermediate matrices of sizes $(n \times m)$ and $(d \times g)$.

Many other algorithms have been proposed: for instance, the sequential algorithm [POD 91], the genetic algorithm [HAN 00, NIE 05], simulated annealing [BRY 05] and tabu search [VON 09]. In [PUO 08] and [TIB 99], the authors

compared these approaches with the use of classical clustering algorithms applied separately on the two sets. Furthermore, in some situations such as in bioinformatics, it is more interesting to divide the data matrix into blocks, which specify a unique partition for objects and a unique partition for variables. Specifically, the data matrix is divided into blocks where, conditionally to each class of objects, a different partition of the variables is allowed. Rocci and Vichi [ROC 08] proposed a generalized double k-means.

I.4.2. *Two-mode hierarchical clustering*

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

I.4.2.1. *Separate clustering*

In a study on the joint use of classification and analysis of correspondence, Jambu [JAM 76] established research links between the two hierarchical classifications obtained from the two sets I and J. The author used time-budgets where I represents a set of population types while J is a set of activity types 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 [GRE 88b] proposed the same approach and provided a simple graphical procedure that is useful in interpreting a significant chi-squared statistic of a contingency table. In a similar way, Camiz and Denimal [CAM 98] analyzed a two-way contingency table using two hierarchies obtained by a classical approach (Ward's method) on each set I and J.

I.4.2.2. *Simultaneous clustering*

Corsten and Denis [COR 90] proposed a two-mode hierarchical clustering. Toledano and Brousse [TOL 77] posed a similar problem: to simultaneously build homogeneous groups of individuals and groups of variables. Their objective is the search for two hierarchies checking this property. For this, they proposed an algorithm, called the double aggregation that seeks the best couple of lines or columns to be incorporated at each iteration. Eckes and Orlik [ECK 93] proposed an agglomerative algorithm for constructing a two-mode hierarchical classification. At each step of the process, the proposed algorithm merges biclusters whose fusion results in the smallest possible increase in an internal heterogeneity measure. This one is based on the variance within the respective cluster and the squared deviation of its mean from the maximum entry in the original data.

For binary data, a two-mode hierarchical clustering algorithm based on dissimilarity measures, derived from the latent block model and illustrated in Figure I.5, was developed by Jollois and Nadif [JOL 04]. Besides, in [JOL 03], a hybrid method jointly using two-mode partitioning and two-mode hierarchical clustering was proposed, enabling us to process large data sets. Both techniques derive from the latent block model described in details in Chapter 1.

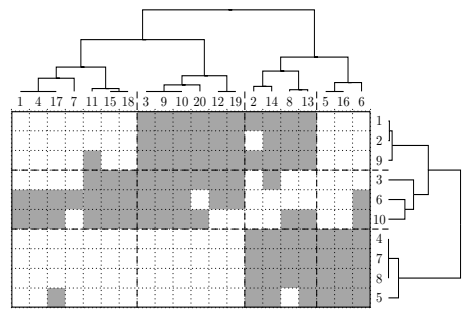

Figure I.5. *Two-mode hierarchical clustering for binary data*

I.4.3. *Direct or block clustering*

In the earliest and most cited biclustering formulation, known as direct or block clustering, Hartigan [HAR 72] defined three families of clusters: the cluster of *responses* or observed values, i.e. a bicluster on the set of $I \times J$, the marginal cluster of objects on I and the marginal cluster of variables (columns) on J. Thus, he suggested the following structures:

– The three-tree structure which requires that all three families of clusters on $I \times J$, I and J are trees.

– The partitioned responses structure which assumes that the clusters of response partition $I \times J$, but that the marginal row clusters and the marginal column clusters form trees on I and J, respectively.

– The three partitions structure which requires that the three families are partitions of $I \times J$, I and J and which is the previous two-mode partitioning.

Note that the observed values must be comparable among variables as well as in the original data or after some previous normalization as discussed in section I.1.3. Then using a stepwise divisive method, Hartigan [HAR 72] developed an algorithm minimizing criterion [I.1]. Furthermore, Tibshirani *et al.* [TIB 99] added a backward pruning and devised a permutation-based method for deciding on the optimal number of blocks. Duffy and Quiroz [DUF 91] proposed another permutation-based algorithm for the same type of structures such that this approach can be extended to a wide variety of data, including matrices of categorical data. Besides, Eckes and Orlik [ECK 93] developed a hierarchical agglomerative algorithm to obtain a hierarchy of blocks.

I.4.4. *Biclustering*

Since the late 1990s, biclustering has been the term most widely used in bioinformatics to identify co-clustering. In this

situation, the structure is no longer limited to three families of clusters as suggested by Hartigan but generalizes the previous co-clustering structure. Its objective is to find a set of K biclusters denoted as $B = \{B_1, \ldots, B_K\}$, where each bicluster B_k is defined as follows:

$$B_k = (I_k, J_k) \text{ with } I_k \subseteq I \text{ and } J_k \subseteq J.$$

Different structures of biclusters exist [MAD 04] and some of them are illustrated in Figure I.6. For instance, S1 corresponds to four biclusters of different sizes with no overlap, S2 corresponds to four biclusters of the same size with overlap on two dimensions, S3 corresponds to four biclusters of the same size with a total overlap on one dimension and S4 corresponds to four biclusters of different sizes with multiple overlaps on one dimension.

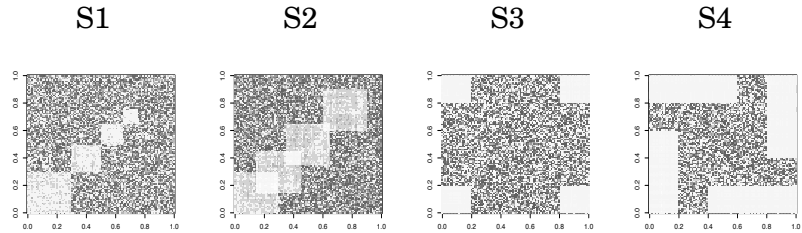

Figure I.6. *Examples of bicluster structures in artificial data sets*

In this context, consider the data matrix x where I is the set of n genes represented by d-dimensional vectors, J is the set of d conditions and x_{ij} represents the expression level of gene i under condition j. A bicluster B_k is a submatrix of x defined by a subset of genes I_k and a subset of conditions J_k. The biclustering aims, for instance, to identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. 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 factor and a multiplicative factor between rows and columns, respectively. Biclusters can also be identified as a combination of these three models. Generally, the searched patterns are based on the correlation among the features or on the coherence among the values in the bicluster as discussed in Madeira's survey [MAD 04]. 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 proposed in the literature and applied to various domains. One of the most active domains is bioinformatics, especially microarray data analysis. The microarrays simultaneously measure the expression of a whole genome under different experimental conditions. Based on these microarray data, we can discover groups of genes with similar expression profiles over a subset of experimental conditions [GOR 02]. The hypothesis is that genes with similar expression profiles should have similar biological functions. The biclustering is also applied to drug activity data in order to associate common properties of chemical compound with common groups of features [LIU 03]. Two of the most popular biclustering methods used in bioinformatics are Chench and Church's algorithm and the plaid model.

Cheng and Church [CHE 00] were the first to propose a biclustering algorithm for microarray data analysis. Their algorithmic framework represents the biclustering problem

as an optimization problem, defining a score for each bicluster candidate and developing heuristics to solve the constrained optimization problem defined by this score function. They consider that biclusters follow an additive model and use a greedy iterative search to minimize the mean square residue (MSR). This algorithm identifies biclusters $B_k = (I_k, J_k)$ one by one by using the score

$$MSR(B_k) = \frac{1}{|I_k| \times |J_k|} \sum_{i,j|i \in I_k, j \in J_k} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2,$$

where $\bar{x}_{i.} = \frac{1}{|I_k|} \sum_{i|i \in I_k} x_{ij}$, $x_{.j} = \frac{1}{|J_k|} \sum_{j|j \in J_k} x_{ij}$ and $\bar{x}_{..} = \frac{1}{|I_k| \times |J_k|} \sum_{i,j|i \in I_k, j \in J_k} x_{ij}$.

Lazzeroni and Owen [LAZ 00] proposed the popular plaid model that was improved by Turner *et al.* [TUR 05]. They assume that biclusters are organized in layers and follow a given statistical model incorporating additive two-way ANOVA models. The values of x_{ij} belonging to a bicluster B_k (layer) depend on four parameters: a constant μ_0 describing the background layer, μ_k the average of B_k, α_{ik} and β_{jk} allowing us to identify, respectively, a subset of genes and examples with identical responses. The values of x_{ij} are therefore represented as

$$x_{ij} = \mu_0 + \sum_{k=1}^{K} \sum_{i,j|i \in I_k, j \in I_k} (\mu_k + \alpha_{ik} + \beta_{jk}).$$

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.

Tanay *et al.* [TAN 02] developed statistical-algorithmic method for bicluster analysis (SAMBA), an approach based

on the graph theory coupled with statistical modeling of the data. Prelic *et al.* [PRE 06] made a comparative study of different biclustering methods for gene expression. They used the Bimax algorithm, a very simple divide and conquer approach as a reference to investigate the usefulness of different biclustering algorithms. They concluded that Bimax produces results similar to those of more complex methods. Note that in such situations, no criterion can be defined and this approach has led to many heuristics: [OYA 01] (ping-pong algorithm), [IHM 02, IHM 04a, IHM 04b, BEN 03, BER 03, TAN 05, TCH 06]. Showing a connection between spectral partition and crossing minimization, Ahmed and Whokhar [AHM 07] developed an efficient biclustering.

Gupta and Aggarwal [GUP 10] introduced the use of mutual information to find relevant biclusters in gene expression data. Mitra and Banka [MIT 06] pointed out that biclustering in microarray data consists of finding sub-matrices, i.e. maximal subgroups of genes and subgroups of conditions where the genes exhibit highly correlated activities over a range of conditions. They have shown that these two objectives are mutually conflicting and they proposed a multiobjective evolutionary biclustering framework by incorporating local search strategies. Cho and Dhillon [CHO 08] proposed a fast biclustering algorithm that minimizes the sum squared residue and generates a checkerboard structure in the data matrix. Erten and Sözdinler [ERT 10] developed their localization procedure that improves the performance of a greedy iterative biclustering algorithm and applied it to microarray data sets. Several other methods were proposed in the literature, a good survey of biclustering methods for biological data analysis was published by Madeira and Oliveira [MAD 04], they enumerated more than 15 algorithms used in this context. Furthermore, to improve the performance of these biclustering methods, Hanczar and Nadif [HAN 11, HAN 12b] adapted the bagging approach to

biclustering problems. The principle consists of generating a set of biclusters and aggregating the results. On simulated and real data sets, this approach appears effective.

I.4.5. *Other structures and other aims*

I.4.5.1. *Block diagonal structure*

In the two-mode partitioning algorithms, constraints can be added to obtain a structure of diagonal blocks after row and column reordering: the row partition and the column partition must have the same number of clusters ($g = m$) and the diagonal biclusters (I_k, J_k) must take a different form 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. Criterion [I.4] with the constraint that the matrix a $= (a_{k\ell})$ is the identity matrix is well adapted to this situation and was used, for instance, by Garcia and Proth [GAR 86] in a group technology application.

This type of approach is also known as *block seriation*. The techniques of seriation, applied in various fields such as sociology, archaeology, botany and zoology, amount to finding a permutation of rows associated with a permutation of columns allowing us to extract the data from a latent order. Block seriation corresponds to a particular approach to this problem and in this context, Marcotorchino [MAR 91a] proposed a solution using linear programming. The BEA [MCC 72, ARA 90] can also be used to obtain a block diagonal structure. Recently, a normalized generalized modularity criterion for binary and categorical data with the aim of co-clustering, was proposed by Labiod and Nadif [LAB 11b]. For its maximization, the authors developed a spectral algorithm and studied the problem of the number of blocks.

Various other approaches have been proposed: Pensa *et al.*, [PEN 05] developed an algorithm leading to a block diagonal

structure with the availability of overlapping. In [DHI 01], modeling the document collection as a bipartite graph between documents and words, the author posed 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.

I.4.5.2. *Different column clustering for each row cluster*

Some authors [POL 02, ROC 08] suggested treating the two-mode clustering by first clustering the rows and then for each row cluster, clustering the columns. For instance, in the partitioning situation, conditionally to each class of row partition, a different partition of columns is allowed.

I.4.5.3. *Multi-way data*

More complex situations exist 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 tables defined on more than two categorical data are examples of array-valued or multi-way data. Some of the previous approaches have been extended to this situation. For instance, Ambroise and Govaert [AMB 02] proposed a clustering of this multiway data along all its dimensions simultaneously using a model-based strategy, and Peng *et al.* [PEN 08] proposed the subspace clustering algorithm.

I.4.5.4. *Principal component analysis and correspondence analysis*

The principal component analysis and its variants such as correspondence analysis methods are also applicable to matrices defined on two sets simultaneously and obtain results on these two sets. It is also possible to show that

certain methods of cross-classification can be seen as principal component analysis under constraints.

I.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 of the data. For most problems, a proper choice of the distance measure can be the main practical difficulty through which much of the arbitrariness of the results can enter. Another class of methods is based on statistical assumptions relating to 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 works of Banfield and Raftery [BAN 93], Celeux and Govaert [CEL 92, CEL 93] and McLachlan [MCL 82] are recent examples, among many others, of this point of view. For co-clustering, even though it is less common and more recent, various models have been proposed.

For instance, Rooth [ROO 95] proposed a probabilistic model for block clustering of contingency tables with a block diagonal structure (see section I.4.5.1 for a description of this structure) and proposed an algorithm that uses formulas similar to the Baum–Welch re-estimation formulas for hidden Markov models. Hartigan [HAR 00] used probabilistic models to perform block clustering on binary data. Nowicki and Snijders [NOW 01] proposed a stochastic block structures model that builds a mixture model for stochastic relationships among objects and identifies the latent cluster via posterior inference. Kemp *et al.* [KEM 06] proposed an

infinite relational model that discovers stochastic structure in relational data in the form of binary observations. Airoldi *et al.* [AIR 08] proposed a mixed membership stochastic block model that relaxes the single-latent-role restriction in stochastic block structure models.

Govaert and Nadif [GOV 03, GOV 08] proposed the latent block model for binary data (binary latent block model) and they extended this model to contingency table (Poisson latent block model) in [GOV 10]. In each case, variational approach, also called a mean-field approximation, of expectation-maximization (EM) algorithm [DEM 77] was used to estimate the parameters and the partitions. Lashkari and Golland [LAS 09] proposed the same generative model and also used a variational approach. They showed that this model has modeling assumptions in common with the Bregman co-clustering of Banerjee *et al.* [BAN 07]. In analyzing continuous data in gene expression context, Jagalur *et al.* [JAG 07] 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 classification EM (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 [TAK 02] proposed a greedy algorithm to estimate the parameters and the two partitions of the latent block model simultaneously (classification approach) and used the Akaike criterion as a stopping rule. In the contingency table situation, Hofmann and Puzicha [HOF 99a] presented different clustering models and in the two-sided clustering situation, the modeling assumptions are 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 a market application, Deodhar and Ghosh [DEO 07] proposed a model-based co-clustering model that can be viewed as an extension of the latent block model, taking into account customer and product attributes. The proposed algorithm interleaves the clustering of customers and products and the construction of prediction models.

Different Bayesian approaches to this kind of model have recently been proposed. Shan and Banerjee [SHA 08] and Dijk *et al.* [DIJ 09] 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. Similar works on the latest block model have been developed by Meeds and Roweis [MEE 07], which have 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, in [POI 08], the authors 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 [KLE 08] proposed a mixture model, and a statistical model of collaborative filtering similar to the Bernoulli latent block model proposed by Ungar and Foster [UNG 98]. For estimating the model parameters, they have developed different methods including variations of the k-means algorithm and Gibbs sampling. Shafiei and Millios [SHA 06] extended these approaches and proposed a model-based overlapping co-clustering able to work with any regular exponential family distribution.

Model-based approaches have also been used to treat the situation described in section I.4.5.2, where different column clusterings are used for each row cluster. In the analysis of gene expression data, Pollard and van der Laan [POL 02] 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 [LI 06] used mixtures of Poisson distribution to model the multivariate distribution of the word counts in the document within each class.

I.6. Outline

In this book, we mainly deal with the two-mode partitioning under different approaches but with particular attention to a probabilistic approach. This book is organized as follows.

– Chapter 1 concerns clustering in general and model-based clustering in particular. We briefly review classical clustering methods and focus on the mixture model. We present and discuss the use of different mixtures adapted to different types of data. The algorithms used are described, and related works with different classical methods are presented and discussed. This chapter will be useful for tackling the problem of co-clustering under the mixture approach.

– Chapter 2 is devoted to the latent block model proposed in the mixture approach context. We discuss this model in detail and show its significance in co-clustering. Various algorithms are presented in a general context.

– Chapter 3 focuses on binary and categorical data. It presents the appropriated latent block mixture models in detail. Variants of these models and of algorithms are presented and illustrated by examples.

– Chapter 4 is devoted to the contingency table. Mutual information, phi-squared coefficient and model-based co-clustering are studied. Models, algorithms and connections among different approaches are described and illustrated.

– Chapter 5 presents the case of continuous data. In the same manner, the different approaches used in the previous chapters are extended to this case.