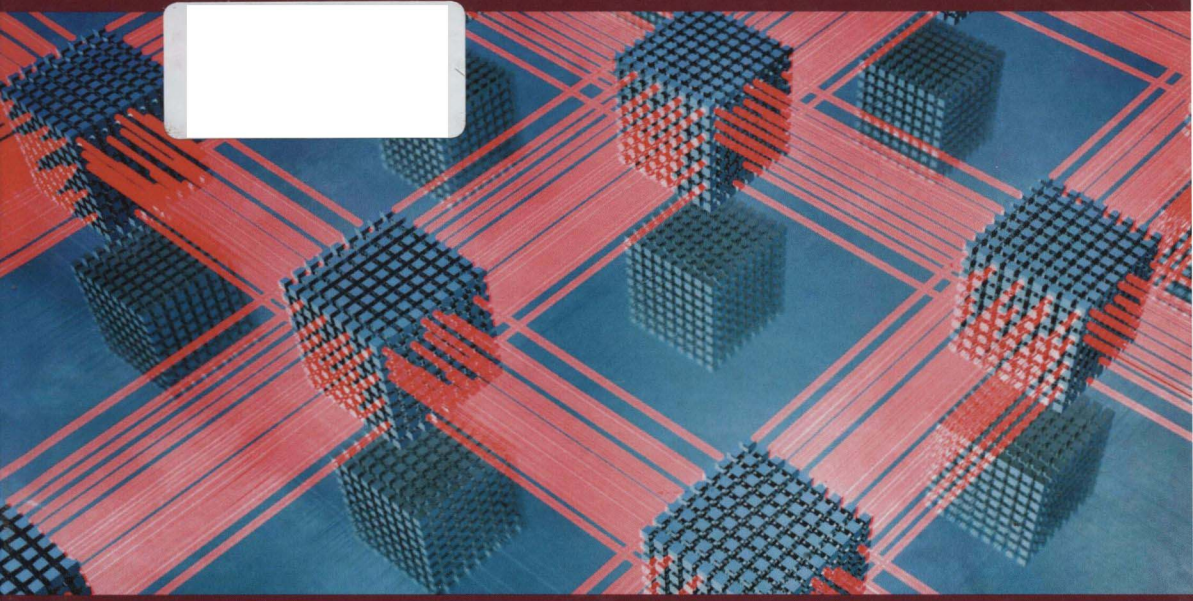
COMPUTER ENGINEERING SERIES



Co-Clustering

Models, Algorithms and Applications

Gérard Govaert
Mohamed Nadif

iSTE WILEY

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

Co-Clustering

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 $\mathrm{x} = (x_{ij})$ 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

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

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 $x_{ij} \in \mathbb{R}$. 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×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

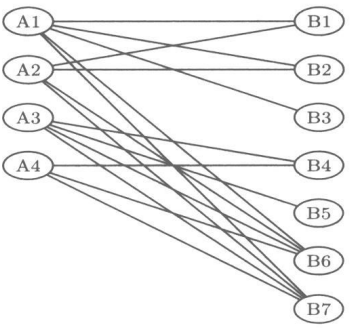


Figure I.3. *Binary data and its associated bipartite graph*

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".