

Behaviormetrics:  
Quantitative Approaches to Human Behavior 5

Tadashi Imaizumi  
Atsuhō Nakayama  
Satoru Yokoyama *Editors*

# Advanced Studies in Behaviormetrics and Data Science

Essays in Honor of Akinori Okada

 Springer

# **Behaviormetrics: Quantitative Approaches to Human Behavior**

Volume 5

## **Series Editor**

Akinori Okada, Professor Emeritus, Rikkyo University, Tokyo, Japan

This series covers in their entirety the elements of behaviormetrics, a term that encompasses all quantitative approaches of research to disclose and understand human behavior in the broadest sense. The term includes the concept, theory, model, algorithm, method, and application of quantitative approaches from theoretical or conceptual studies to empirical or practical application studies to comprehend human behavior. The Behaviormetrics series deals with a wide range of topics of data analysis and of developing new models, algorithms, and methods to analyze these data.

The characteristics featured in the series have four aspects. The first is the variety of the methods utilized in data analysis and a newly developed method that includes not only standard or general statistical methods or psychometric methods traditionally used in data analysis, but also includes cluster analysis, multidimensional scaling, machine learning, corresponding analysis, biplot, network analysis and graph theory, conjoint measurement, biclustering, visualization, and data and web mining. The second aspect is the variety of types of data including ranking, categorical, preference, functional, angle, contextual, nominal, multi-mode multi-way, contextual, continuous, discrete, high-dimensional, and sparse data. The third comprises the varied procedures by which the data are collected: by survey, experiment, sensor devices, and purchase records, and other means. The fourth aspect of the Behaviormetrics series is the diversity of fields from which the data are derived, including marketing and consumer behavior, sociology, psychology, education, archaeology, medicine, economics, political and policy science, cognitive science, public administration, pharmacy, engineering, urban planning, agriculture and forestry science, and brain science.

In essence, the purpose of this series is to describe the new horizons opening up in behaviormetrics—approaches to understanding and disclosing human behaviors both in the analyses of diverse data by a wide range of methods and in the development of new methods to analyze these data.

### **Editor in Chief**

Akinori Okada (Rikkyo University)

### **Managing Editor**

Daniel Baier (University of Bayreuth)

Giuseppe Bove (Roma Tre University)

Takahiro Hoshino (Keio University)

More information about this series at <http://www.springer.com/series/16001>

Tadashi Imaizumi · Atsuhō Nakayama ·  
Satoru Yokoyama  
Editors

# Advanced Studies in Behaviormetrics and Data Science

Essays in Honor of Akinori Okada

 Springer

*Editors*

Tadashi Imaizumi  
School of Management  
and Information Sciences  
Tama University  
Tokyo, Japan

Atsuho Nakayama  
Graduate School of Management  
Tokyo Metropolitan University  
Tokyo, Japan

Satoru Yokoyama  
School of Business  
Aoyama Gakuin University  
Tokyo, Japan

ISSN 2524-4027

ISSN 2524-4035 (electronic)

Behaviormetrics: Quantitative Approaches to Human Behavior

ISBN 978-981-15-2699-2

ISBN 978-981-15-2700-5 (eBook)

<https://doi.org/10.1007/978-981-15-2700-5>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Foreword

I am very pleased to write a foreword for the Festschrift of Prof. Akinori Okada. He was born in Tokyo in September 1943, and all his academic life has been active. Now he is 77-years old; it is the age of celebration, according to the traditional Japanese age system.

Professor Okada obtained a job at Chiba University in 1971 and moved to Rikkyo University in the next year to be a Senior Lecturer and then Associate Professor of statistics and operations research. He received his Ph.D. in Engineering from Keio University in 1979. He spent 35 years in Rikkyo University until he further moved to Tama University in 2007, where his best collaborator Prof. Tadashi Imaizumi serves.

Professor Okada has made significant contributions to the fields of multidimensional scaling and cluster analysis, especially the analysis of asymmetric relationships. His works on these fields are so influential that he edited more than 15 books in Japanese and English on multidimensional scaling, cluster analysis, statistics, linear algebra, operations research, data analysis, and marketing science. The book entitled ‘Operations Research: Introduction to Management Science’ [Operêshonzu Risâchi–Keieikagaku Nyumon–] written with Dr. Ken-ichi Goto, published in 1987, has been used as a standard textbook of operations research for more than 30 years in courses in the departments of social science schools in Japan.

I share a good memory with Prof. Okada; we hosted together the first International Meeting of the Psychometric Society (IMPS) in Osaka in Japan in 2001. Prior to that, Prof. Okada and the late Prof. Haruo Yanai participated in the council meeting of the Psychometric Society in the U.S. One of the main purposes for the visit was to propose we should host an annual meeting in Asia, particularly in Japan for the first time. As a result, the meeting was successfully held in July 2001, where Prof. Okada was the Vice President of the executive committee and I served as the chair of the local organizing committee. The meeting was a great success; more than 400 participants gathered and accordingly, the financial results were in a sound condition. The T-shirts with the IMPS logo shown by Prof. Heiser in the photo below was a gift for Prof. Jacqueline Meulman, as a small token of the members’ appreciation; although she had seriously prepared for the meeting together with us, her sickness prevented her from attending the meeting in the end.



**Photo 1** Professors Akinori Okada, Haruo Yanai, Willem Heiser, Kazuo Shigemasa and Yutaka Kano from the flush left. Photo taken on July 19, 2001, just after the IMPS2001

Professor Okada is the series editor in Springer named *Behaviormetrics: Quantitative Approaches to Human Behavior*. He launched the joint meeting of the German Classification Society and the Japanese Classification Society in corporation with German colleagues, and also organized the joint meeting with the Classification and Data Analysis Group of the Italian Statistical Society and the Japanese Classification Society in corporation with Italian colleagues. In addition to this, Prof. Okada served as the President of the International Federation of Classification Societies (2016–2017).

He led the Behaviormetric Society as the President (2012–2015), and has still been a member of the board of directors of the Society; his achievements have proven that he is one of the distinguished leaders of the behaviormetrics and classification societies.

In Japan, they say we should make a life plan for up to hundred years now, since the average life span continues to grow. Professor Okada still has a quarter of a century more of his life. We hope that he can continue to enjoy his life and remain active as an academician.

Osaka, Japan  
November 2019

Yutaka Kano

# Preface

The year 2019 is the 40th anniversary of the doctorate degree of Akinori Okada, who is a Professor Emeritus at Rikkyo University, Japan. We are delighted to work as Editors for this 'Festschrift' of him. He received his Ph.D. in Engineering from Keio University, Tokyo in 1979. During the course of his long and distinguished career of more than 45 years, he has made significant research contributions to theory and applications in areas such as multidimensional scaling, cluster analysis, psychometrics, data analysis, operations research, data science, marketing research, consumer behavior, psychological, and social human relationships. Today these research contributions are essential constituents of the realm of 'behaviormetrics'. Akinori Okada has played an important role as a leader in developing behaviormetrics. Especially his research contribution toward the development and application of asymmetric multidimensional scaling and cluster analysis are significant. His work has been published in journals such as *Advances in Data Analysis and Classification*, *Behaviormetrika*, *Journal of Classification*, *Psychonomic Science*, *Journal of Applied Psychology*, *Organizational Behavior and Human Performance*, *Japanese Journal of Behaviormetrics* [Kôdo Keiryogaku], *Japanese Psychological Review* [Shinrigaku Hyôron], *Sociological Theory and Methods* [Riron to Hôhô], *Communications of the Operations Research* [Operêshionzu Risâchi], *Japanese Journal of Applied Statistics* [Ôyotôkeigaku], *Japanese Review of Clinical Ophthalmology* [Ganka Rinsho Ihô], *The Journal of the Institute of Electronics and Communication Engineers of Japan* [Denshi Jyôhō Tsushin Gakkaishi Shi] among other things, as well as in numerous refereed proceedings volumes.

He also played an important role as a scientific leader in behaviormetrics, and especially contributed his effort to introduce the quantitative concept in social sciences. He is the series editor of 'Behaviormetrics: Quantitative Approaches to Human Behavior', published by Springer, which covers all aspects of Behaviormetrics; theory, concept, method, and application in order to disclose and understand human behavior. He is one of the founding managing editors of the 'Advances in Data Analysis and Classification'. He also is the founding editor of the 'Bulletin of Data Analysis of Japanese Classification Society' or 'Dêta Bunseki no Riron to Ôyo' in Japanese. Akinori Okada was the President of the



International Federation of Classification Societies (2016–2017), the Behaviormetric Society (2012–2015), Japanese Classification Society (2005–2009), and chaired the program committee of numerous international conferences. He is a Research Fellow of the Operations Research Society of Japan. He has been an outside director of SHL-Japan Limited since 2002 as a statistician. Akinori Okada is a great mentor of many students. Three of his former students decided to honor him for his outstanding achievements in behaviormetrics and data science by inviting his colleagues and friends to contribute articles for this ‘Festschrift’, who sent us articles of high quality to us. Two of us wrote an article each as well. The present ‘Festschrift’ focuses on the latest developments in behaviormetrics and data science, and covers both theoretical aspect and applications to a wide range of areas including psychology, marketing science, sociology, social survey, operations research, etc. The contributions to this volume are intended for researchers and practitioners who are interested in the latest developments and applications in these fields. The present volume consists of two parts which express two aspects of the research of behaviormetrics and data science by Akinori Okada: a theoretically-oriented part and an application-oriented part. Contributions are ordered alphabetically based on the corresponding authors’ names within each of these two parts. We have to confess that there are several anomalies in the order of contributions due to the inattention of the first editor. We would like to express our deepest appreciation to authors for their contributions to the volume and cooperation while we edited the volume. We want to show our heartiest gratitude to Mr. Reginald Williams and Ms. Yasuko Hase for thoughtfully helping us in English for the e-mail of inviting authors and of the reminder to the authors. We cordially appreciate Mr. Yutaka Hirachi and Ms. Sridevi Purushothaman at Springer Nature for their assistance for publishing the present volume.

Tokyo, Japan  
March 2020

Tadashi Imaizumi  
Atsuhō Nakayama  
Satoru Yokoyama

# Contents

## Part I Theoretically-Oriented

<b>Co-Clustering for Object by Variable Data Matrices</b> . . . . .	3
Hans-Hermann Bock	
<b>How to Use the Hermitian Form Model for Asymmetric MDS</b> . . . . .	19
Naohito Chino	
<b>Asymmetric Scaling Models for Square Contingency Tables: Points, Circles, Arrows and Odds Ratios</b> . . . . .	43
Mark de Rooij	
<b>Flight Passenger Behavior and Airline Fleet Assignment</b> . . . . .	63
Wolfgang Gaul and Christoph Winkler	
<b>Comparing Partitions of the Petersen Graph</b> . . . . .	83
Andreas Geyer-Schulz and Fabian Ball	
<b>Minkowski Distances and Standardisation for Clustering and Classification on High-Dimensional Data</b> . . . . .	103
Christian Hennig	
<b>On Detection of the Unique Dimensions of Asymmetry in Proximity Data</b> . . . . .	119
Tadashi Imaizumi	
<b>Multiple Regression Analysis from Data Science Perspective</b> . . . . .	131
Manabu Iwasaki	
<b>Multiway Extensions of the SVD</b> . . . . .	141
Pieter M. Kroonenberg	
<b>Seriation and Matrix Reordering Methods for Asymmetric One-Mode Two-Way Datasets</b> . . . . .	159
Innar Liiv and Leo Vohandu	

<b>Parsimonious Mixtures of Matrix Variate Bilinear Factor Analyzers</b> . . . . .	177
Michael P. B. Gallaughier and Paul D. McNicholas	
<b>Interval-Valued Scaling of Successive Categories</b> . . . . .	197
Hisao Miyano and Eric J. Beh	
<b>Orthonormal Principal Component Analysis for Categorical Data as a Transformation of Multiple Correspondence Analysis</b> . . . . .	211
Takashi Murakami	
<b>Identifying Groups With Different Traits Using Fourteen Domains of Social Consciousness: A Multidimensional Latent Class Graded Item Response Theory Model</b> . . . . .	233
Miki Nakai and Fulvia Pennoni	
<b>Quantification Theory: Categories, Variables and Modal Analysis</b> . . . . .	253
Shizuhiko Nishisato	
<b>Clustering via Ant Colonies: Parameter Analysis and Improvement of the Algorithm</b> . . . . .	265
Jeffry Chavarria-Molina, Juan José Fallas-Monge and Javier Trejos-Zelaya	
<b>PowerCA: A Fast Iterative Implementation of Correspondence Analysis</b> . . . . .	283
Alfonso Iodice D’Enza, P. J. F. Groenen and M. Van de Velden	
<b>Modeling Asymmetric Exchanges Between Clusters</b> . . . . .	297
Donatella Vicari	
<b>Exploring Hierarchical Concepts: Theoretical and Application Comparisons</b> . . . . .	315
Carlo Cavicchia, Maurizio Vichi and Giorgia Zaccaria	
<b>Improving Algorithm for Overlapping Cluster Analysis</b> . . . . .	329
Satoru Yokoyama	
<b>Part II Application-Oriented</b>	
<b>Increasing Conversion Rates Through Eye Tracking, TAM, A/B Tests: A Case Study</b> . . . . .	341
Daniel Baier and Alexandra Rese	
<b>Descriptive Analyses of Interrater Agreement for Ordinal Rating Scales</b> . . . . .	355
Giuseppe Bove and Alessio Serafini	
<b>The Globality of Brands—A Question of Methods?</b> . . . . .	367
Michael Löffler and Reinhold Decker	

**Mapping Networks and Trees with Multidimensional Scaling of Proximities** ..... 385  
Willem J. Heiser, Frank M. T. A. Busing and Jacqueline J. Meulman

**Pitfalls in the Construction of Response Scales in Cross-Cultural Surveys: An Example from East Asian Social Survey** ..... 409  
Noriko Iwai and Satomi Yoshino

**Japanese Women’s Attitudes Toward Childrearing: Text Analysis and Multidimensional Scaling** ..... 423  
Kunihiro Kimura

**Consensus or Dissensus in Occupational Prestige Evaluation: A New Approach to Measuring Consensus and Inter-group Variations** ..... 439  
Keiko Nakao

**People and Trust** ..... 453  
Ryozo Yoshino

# Contributors

**Daniel Baier** Chair of Marketing and Innovation, Universitaetsstrasse 30, University of Bayreuth, Bayreuth, Germany

**Fabian Ball** Information Services and Electronic Markets, Institute of Information Systems and Marketing, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**Eric J. Beh** University of Newcastle, Callaghan, NSW, Australia

**Hans-Hermann Bock** Institute of Statistics, RWTH Aachen University, Aachen, Germany

**Giuseppe Bove** Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre, Rome, Italy

**Frank M. T. A. Busing** Faculty of Social and Behavioral Sciences, Leiden University, Leiden, The Netherlands

**Carlo Cavicchia** University of Rome La Sapienza, Rome, Italy

**Jeffry Chavarría-Molina** School of Mathematics, Costa Rica Institute of Technology, Cartago, Costa Rica

**Naohito Chino** Aichi Gakuin University, Nagoya, Japan

**Reinhold Decker** Department of Business Administration and Economics, Bielefeld University, Bielefeld, Germany

**Mark de Rooij** Department of Methodology and Statistics, Institute of Psychology, Leiden University, Leiden, The Netherlands

**Juan José Fallas-Monge** School of Mathematics, Costa Rica Institute of Technology, Cartago, Costa Rica

**Michael P. B. Gallagher** Department of Mathematics and Statistics, McMaster University, Hamilton, Canada

**Wolfgang Gaul** Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**Andreas Geyer-Schulz** Information Services and Electronic Markets, Institute of Information Systems and Marketing, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**P. J. F. Groenen** Econometric Institute, Erasmus University, Rotterdam, The Netherlands

**Willem J. Heiser** Faculty of Social and Behavioral Sciences and Mathematical Institute, Leiden University, Leiden, The Netherlands

**Christian Hennig** Dipartimento di Scienze Statistiche “Paolo Fortunati”, Bologna, Italy

**Tadashi Imaizumi** Tama University, Tama-shi, Tokyo, Japan

**Alfonso Iodice D’Enza** Department of Political Sciences, Università degli Studi di Napoli Federico II, Naples, Italy

**Noriko Iwai** JGSS Research Center, Osaka University of Commerce, Higashi-Osaka, Japan

**Manabu Iwasaki** Yokohama City University, Yokohama, Japan

**Kunihiro Kimura** Tohoku University, Sendai, Japan

**Pieter M. Kroonenberg** Faculty of Social and Behavioural Sciences, Leiden University, Leiden, The Netherlands;  
The Three-Mode Company, Leiden, The Netherlands

**Innar Liiv** Tallinn University of Technology, Tallinn, Estonia

**Michael Löffler** Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany

**Paul D. McNicholas** Department of Mathematics and Statistics, McMaster University, Hamilton, Canada

**Jacqueline J. Meulman** Mathematical Institute, Leiden University, Leiden, The Netherlands;  
Department of Statistics, Stanford University, Stanford, CA, USA

**Hisao Miyano** Chiba University, Chiba, Japan

**Takashi Murakami** Chukyo University, Nagoya, Japan

**Miki Nakai** Department of Social Sciences, College of Social Sciences, Ritsumeikan University, Kyoto, Japan

**Keiko Nakao** Tokyo Metropolitan University, Tokyo, Japan

**Shizuhiko Nishisato** University of Toronto, Toronto, Canada

**Fulvia Pennoni** Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

**Alexandra Rese** Chair of Marketing and Innovation, Universitaetsstrasse 30, University of Bayreuth, Bayreuth, Germany

**Alessio Serafini** Dipartimento di Economia, Università degli Studi di Perugia, Perugia, Italy

**Javier Trejos-Zelaya** CIMPA–School of Mathematics, University of Costa Rica, San José, Costa Rica

**M. Van de Velden** Econometric Institute, Erasmus University, Rotterdam, The Netherlands

**Donatella Vicari** Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy

**Maurizio Vichi** University of Rome La Sapienza, Rome, Italy

**Leo Vohandu** Tallinn University of Technology, Tallinn, Estonia

**Christoph Winkler** Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**Satoru Yokoyama** Department of Marketing, School of Business, Aoyama Gakuin University, Tokyo, Japan

**Ryozo Yoshino** Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan

**Satomi Yoshino** JGSS Research Center, Osaka University of Commerce, Higashi-Osaka, Japan

**Giorgia Zaccaria** University of Rome La Sapienza, Rome, Italy

**Part I**  
**Theoretically-Oriented**



# Co-Clustering for Object by Variable Data Matrices



Hans-Hermann Bock

**Abstract** Co-clustering means the simultaneous clustering of the rows and columns of a two-dimensional data table (biclustering, two-way clustering), in contrast to separately clustering the rows and the columns. Practical applications may be met, e.g., in economics, social sciences, bioinformatics, etc. Various co-clustering models, criteria, and algorithms have been proposed that differ with respect to the considered data types (real-valued, integers, binary data, contingency tables), and also the meaning of rows and columns (samples, variables, factors, time,...). This paper concentrates on the case where rows correspond to (independent) samples or objects, and columns to (typically dependent) variables. We emphasize that here, in general, different similarity or homogeneity concepts must be used for rows and columns. We propose two probabilistic co-clustering approaches: a situation where clusters of objects and of variables refer to two different distribution parameters, and a situation where clusters of ‘highly correlated’ variables (by regression to a latent class-specific factor) are crossed with object clusters that are distinguished by additive effects only. We emphasize here the classical ‘classification approach’, where maximum likelihood criteria are optimized by generalized alternating  $k$ -means type algorithms.

## 1 Co-Clustering

Clustering methods are well-known tools for analyzing and structuring data, intensively investigated in statistics, machine learning and data science, and broadly used in many application domains such as as market and consumer research, psychology and social sciences, microbiology and bioinformatics. The basic problem consists in grouping a given set of objects into homogeneous classes (clusters) on the basis of empirical data that allow to quantify the ‘similarity’ or ‘dissimilarity’ of the objects and so to define the homogeneity within, or the separation between, the classes. In the most simple case there is an  $n \times p$  data matrix  $X = (x_{ij})$ , where values  $x_{ij}$

---

H.-H. Bock (✉)

Institute of Statistics, RWTH Aachen University, Aachen, Germany  
e-mail: [bock@stochastik.rwth-aachen.de](mailto:bock@stochastik.rwth-aachen.de)

© Springer Nature Singapore Pte Ltd. 2020

T. Imaizumi et al. (eds.), *Advanced Studies in Behaviormetrics and Data Science*,  
Behaviormetrics: Quantitative Approaches to Human Behavior 5,  
[https://doi.org/10.1007/978-981-15-2700-5\\_1](https://doi.org/10.1007/978-981-15-2700-5_1)

are recorded for  $n$  objects and  $p$  variables and we look for an appropriate partition  $\mathcal{A} = (A_1, \dots, A_k)$  of the set of objects  $\mathcal{O} = \{1, \dots, n\}$  (the rows of  $X$ ) with classes  $A_1, \dots, A_k$  such that similar row vectors (objects) are united in the same class while row vectors from different classes are hopefully quite dissimilar. Depending on the context, the detected or constructed clusters will be interpreted as (personality) types, consumer groups, music styles, families of plants, gene clusters, etc. Since the early 1960s when clustering methods came up, a large variety of clustering models and clustering algorithms have been developed for different data types, see, e.g., Bock [7, 9], Jain and Dubes [26], Miyamoto, Ichihashi, and Honda [32], McLachlan and Krishnan [31], Basu, Davidson, and Wagstaff [5], Aggarwal and Reddy [1], and Hennig, Meila, Murtagh, and Rocci [25].

Co-clustering (biclustering, two-way clustering, block clustering) means the simultaneous (i.e., not separate) clustering of the rows and columns of a data matrix by determining an appropriate partition  $\mathcal{A} = (A_1, \dots, A_k)$  of the rows *together* with an appropriate partition  $\mathcal{B} = (B_1, \dots, B_\ell)$  of the set of columns  $\mathcal{M} = \{1, \dots, p\}$  such that both row and column clusters are ‘homogeneous’ and reflect the hidden interplay between row and column effects. Biclustering provides an aggregated view on the similarity structure within the sets  $\mathcal{O}$  and  $\mathcal{M}$  of objects and columns, respectively, and also can serve in order to reduce a large data table with  $n \cdot p$  entries to a manageable size with only  $k \cdot \ell$  blocks  $A_k \times B_\ell$  together with their characterizations (data compression). Often the aggregated view on the blocks will provide a better insight into the latent relationships and interactions that may exist between objects and variables than a detailed analysis of the numerous entries  $x_{ij}$ . Many applications underline the usefulness of co-clustering methods, e.g., in marketing (Arabie, Schleutermann, Daws, & Hubert [3]; Gaul & Schader [18]; Baier, Gaul, & Schader [4]), psychology and social sciences (Kiers, Vicari, & Vichi [27]; Schepers, Bock, & Van Mechelen [38]), bioinformatics and gene analysis (Cheng & Church [16]; Madeira & Oliveira [28]; Turner, Bailey, Krzanowski, & Hemmingway [39]; Alfò, Martella, & Vichi [2]; Martella, Alfò, & Vichi [30]; Cho & Dhillon [17]; Martella & Vichi [29]; Pontes, Giráldez, & Aguilar-Ruiz [33]), and text mining (Dhillon [19]).

A range of co-clustering methods have been proposed in the past, see, e.g., the surveys in Van Mechelen, Bock, and De Boeck [40], Madeira and Oliveira [28], Charrad and Ben Ahmed [14], Govaert and Nadif [23, 24] and methodological articles such as Bock [8, 10–12], Govaert [20], Vichi [41], Govaert and Nadif [21, 22, 24], Rocci and Vichi [34], Salah and Nadif [35], Schepers, Bock, and Van Mechelen [38] and Schepers and Hofmans [36]. These methods differ, e.g.,

- By the type of observed data values  $x_{ij}$ , e.g., real-valued, integer, categorical, binary, mixed, etc.
- by the meaning of the entries  $x_{ij}$ , e.g., association values, measurements, frequencies, etc.
- by the classification structure, e.g., hierarchical versus nonhierarchical, hard versus fuzzy classifications, and mixtures.

- by the modeling approach using, e.g., probabilistic models, empirical concepts, optimization criteria, and algorithms, etc.
- by the practical meaning of the rows and columns.

Concerning this latter issue, we may distinguish between cases where rows and columns denote the categories of two given nominal factors (e.g., the crop variety  $i$  with the fertilizer  $j$  yields  $x_{ij}$  tons of cereals), and cases of the object  $\times$  variable type mentioned in the first paragraph above (e.g., object  $i$  realizes the value  $x_{ij}$  for variable  $j$ ). While the two-factor case is typically symmetric insofar as clustering of both rows and columns is (or may be) based on the nearness of corresponding entries in the rows and columns, respectively, this may be misleading in the second unsymmetric case since, differently from the objects (rows), the similarity of variables (columns) is typically expressed in terms of mutual dependencies or interrelationships. Insofar, in the object  $\times$  variable case, clustering of rows and columns should typically be based on different distance or similarity indices that must be integrated into a joint two-way clustering model.

In this paper, we consider two situations of this latter type and provide, as a paradigm for more complex situations, suitable probabilistic co-clustering models and corresponding  $k$ -means type algorithms: In Sect. 2 we describe a two-way two-parameter biclustering model where the row partition  $\mathcal{A}$  refers to the first parameter (cluster means) while the column partition  $\mathcal{B}$  is induced by the values of the second one (class-specific variances). A more sophisticated and novel co-clustering model is described in Sect. 4, where object classes are characterized by a class-specific *mean value* (main effect) while additionally each class of variables is characterized by a class-specific *latent factor* that is estimated together with the column partition. As a prelude for this latter two-way model we consider in Sect. 3 a (one-way) clustering algorithm for variables only, proposed by Vigneau and Qannari [43] that is related to correlation and latent factor concepts, and show that it can be derived from a probabilistic one-way clustering model. In Sect. 4 this model will be integrated in the two-way clustering case. Section 5 concludes with some remarks and possible extensions.

## 2 Co-clustering with Class-Specific Variances in the Variable Clusters

We have emphasized in Sect. 1 that for an object  $\times$  variable matrix  $X = (x_{ij})$ , clustering of variables (columns of  $X$ ) may be inspired by other purposes or characterizations than when clustering objects (rows of  $X$ ). In this section, we consider a simple example for such a co-clustering problem and describe a model where object clusters are characterized by cluster means (main effects) while clusters of variables are distinguished by different variability of the data. More specifically, we consider the following probabilistic co-clustering model for independent normally distributed random variables  $X_{ij}$ :

$$X_{ij} \sim \mathcal{N}(\mu_s, \sigma_t^2) \quad \text{for } i \in A_s, j \in B_t, s = 1, \dots, k, t = 1, \dots, \ell \quad (1)$$

with the  $k$ -partition  $\mathcal{A} = (A_1, \dots, A_k)$  of the  $n$  rows, the  $\ell$ -partition  $\mathcal{B} = (B_1, \dots, B_\ell)$  of the  $p$  columns of the matrix  $X = (X_{ij})$ , where row clusters  $A_s$  are characterized by cluster-specific expectations  $\mu_s$  while column classes  $B_t$  are characterized by class-specific variances  $\sigma_t^2$ . In this situation, maximum likelihood estimation of the unknown parameters  $\mathcal{A}, \mathcal{B}, \mu = (\mu_1, \dots, \mu_k)$ , and  $\sigma = (\sigma_1^2, \dots, \sigma_\ell^2)$  (for fixed  $k$  and  $\ell$ ) is equivalent to the minimization of the co-clustering criterion

$$Q(\mathcal{A}, \mathcal{B}, \mu, \sigma; X) := \sum_{s=1}^k \sum_{t=1}^{\ell} \sum_{i \in A_s} \sum_{j \in B_t} \left[ \frac{(x_{ij} - \mu_s)^2}{\sigma_t^2} + \log \sigma_t^2 \right] \rightarrow \min_{\mathcal{A}, \mathcal{B}, \mu, \sigma}. \quad (2)$$

Equating to zero the partial derivatives w.r.t.  $\mu_s$  and  $\sigma_t^2$  yields the (implicit) formulas for the estimates  $\hat{\mu}_s$  and  $\hat{\sigma}_t^2$ :

$$\mu_s = \left[ \sum_{t=1}^{\ell} \frac{|B_t|}{\sigma_t^2} \bar{x}_{A_s, B_t} \right] / \left[ \sum_{t=1}^{\ell} \frac{|B_t|}{\sigma_t^2} \right] \quad (3)$$

$$\begin{aligned} \sigma_t^2 &= \frac{1}{n \cdot |B_t|} \cdot \sum_{j \in B_t} \sum_{s=1}^k \sum_{i \in A_s} (x_{ij} - \mu_s)^2 \\ &= \frac{1}{n \cdot |B_t|} \cdot \sum_{j \in B_t} \sum_{s=1}^k \left[ \sum_{i \in A_s} (x_{ij} - \bar{x}_{A_s, j})^2 + |A_s| \cdot (\bar{x}_{A_s, j} - \mu_s)^2 \right]. \end{aligned} \quad (4)$$

Here  $|A_s|, |B_t|$  are the class sizes, and we use largely self-explanatory notations such as

$$\begin{aligned} \bar{x}_{A_s, j} &:= \sum_{i \in A_s} x_{ij} / |A_s|, & \bar{x}_{i, B_t} &:= \sum_{j \in B_t} x_{ij} / |B_t| \\ \bar{x}_{A_s, B_t} &:= \sum_{i \in A_s} \sum_{j \in B_t} x_{ij} / (|A_s| \cdot |B_t|), & \bar{x}_{\cdot, \cdot} &:= \sum_{i=1}^n \sum_{j=1}^n x_{ij} / (n \cdot p). \end{aligned}$$

So the estimate  $\hat{\mu}_s = \mu_s$  is a weighted mean of the  $\ell$  block means  $\bar{x}_{A_s, B_t}$  (with weights inversely proportional to  $\sigma_t^2 / |B_t|$ , the variance of the mean  $\bar{X}_{i, B_t}$  in the class  $B_t$ ) and the estimate  $\hat{\sigma}_t^2 = \sigma_t^2$  comprises terms that measure the variability within  $A_s$  (for the variables  $j \in B_t$ ) and the distance between the individual means  $\bar{x}_{A_s, j}$  from the class-specific estimated expectations  $\mu_s$ .

Since it is impossible to obtain explicit formulas for both estimates we propose to resolve the co-clustering problem (2) by the following iterative algorithm of the  $k$ -means type:

1. Begin with two initial partitions  $\mathcal{A}$  and  $\mathcal{B}$  and an initial estimate for  $\sigma$  (e.g., with  $\sigma_t^2$  the empirical variance of the data values in the  $|B_t|$  columns of  $X$  corresponding to  $B_t$ );

2. Estimate the object-class-specific expectations  $\mu_s$  by (3) (i.e., minimize  $Q$  w.r.t.  $\mu$ );
3. Estimate the variable-class-specific variances  $\sigma_t^2$  by (4) (i.e., minimize  $Q$  w.r.t.  $\sigma$ );
4. For given  $\mathcal{B}$ ,  $\mu$ , and  $\sigma$  minimize  $Q$  w.r.t. the  $k$ -partition  $\mathcal{A}$  of the set of objects  $\mathcal{O} = \{1, \dots, n\}$ . An elementary argumentation shows that the minimum is obtained by the *generalized minimum-distance partition*  $\tilde{\mathcal{A}}$  with object (row) clusters

$$\tilde{\mathcal{A}}_s := \{ i \in \mathcal{O} \mid s = \operatorname{argmin}_{s'=1, \dots, k} d(i, \mu_{s'} | \mathcal{B}, \sigma) \} \quad s = 1, \dots, k,$$

where the distance  $d$  is defined by

$$d(i, \mu_{s'} | \mathcal{B}, \sigma) := \sum_{t=1}^{\ell} \sum_{j \in \mathcal{B}_t} (x_{ij} - \mu_{s'})^2 / \sigma_t^2.$$

5. Update the parameter estimates  $\mu$ ,  $\sigma$  by repeating Steps 2. and 3. for the current partitions  $\tilde{\mathcal{A}}$  and  $\mathcal{B}$ .
6. Given  $\tilde{\mathcal{A}}$ ,  $\mu$ , and  $\sigma$ , minimize  $Q$  w.r.t. the  $\ell$ -partition  $\mathcal{B}$  of the set of variables  $\mathcal{M} = \{1, \dots, p\}$ ; the solution is given by the *generalized minimum-distance partition*  $\tilde{\mathcal{B}}$  with variable (column) clusters

$$\tilde{\mathcal{B}}_t := \{ j \in \mathcal{M} \mid t = \operatorname{argmin}_{t'=1, \dots, \ell} \delta(j, \sigma_{t'}^2 | \tilde{\mathcal{A}}, \mu) \} \quad t = 1, \dots, \ell,$$

where the distance  $\delta$  is defined by

$$\delta(j, \sigma_{t'}^2 | \tilde{\mathcal{A}}, \mu) := \sum_{s=1}^k \sum_{i \in \tilde{\mathcal{A}}_s} (x_{ij} - \mu_s)^2 / \sigma_{t'}^2 + n \cdot \log \sigma_{t'}^2.$$

7. Iterate 2. to 6. until convergence.

Obviously this algorithm decreases successively the criterion  $Q$ , (2), and insofar approximates a solution to the stated co-clustering problem. Note that ties, empty classes, local optima, and oscillating partitions may be possible and must be considered or avoided in a corresponding computer program.

### 3 Clustering of Variables Around Latent Factors

In this section, we describe a method for one-way clustering of the  $p$  variables (columns) of a data matrix  $X = (x_{ij})$  that has been proposed by Vigneau and Qannari [43] and uses squared correlations for measuring the similarity between two variables. In fact, we show that this method and the related clustering criterion can be derived, as a special case, from a relatively general probabilistic clustering

model that characterizes each cluster of variables by a class-specific latent factor. In the following Sect. 4 this model will be integrated in our co-clustering models (12) and (13) for the objects and variables of  $X$ .

In many practical contexts, the similarity of two random variables  $Y_j, Y_{j'}$  is measured by their squared correlation  $r^2(Y_j, Y_{j'}) := \text{Corr}^2(Y_j, Y_{j'})$ . Similarly, in case of a  $n \times p$  data matrix  $X = (x_{ij}) = (y_1, \dots, y_p)$ , where the  $j$ -th column  $y_j = (x_{1j}, \dots, x_{nj})^\top$  represents the  $j$ -th variable, the similarity of  $y_j$  and  $y_{j'}$  (or  $j$  and  $j'$ ) is measured by the square of the empirical correlation

$$r(y_j, y_{j'}) := \frac{s_{y_j, y_{j'}}}{\sqrt{s_{y_j, y_j} s_{y_{j'}, y_{j'}}}}$$

with

$$s_{y_j, y_{j'}} := (1/n) \sum_{i=1}^n (x_{i,j} - \bar{x}_{\cdot,j})(x_{i,j'} - \bar{x}_{\cdot,j'}) = (1/n) y_j^\top y_{j'},$$

where  $\bar{x}_{\cdot,j} := (\sum_{i=1}^n x_{ij})/n$  is the mean of the  $n$  entries in the column  $j$  of  $X$  and the last equality sign holds for centered columns  $y_j$  (i.e.,  $\bar{x}_{\cdot,j} = 0$ ).

Vigneau and Qannari have integrated this similarity concept into the search for an optimal  $\ell$ -partition  $\mathcal{B} = (B_1, \dots, B_\ell)$  of the set  $\mathcal{M}$  of variables (columns of  $X$ ). In order to formulate a corresponding clustering criterion, they define, for each class  $B_t$ , a suitable ‘prototype variable’ or ‘class representative’. Instead of choosing one of the observed variables (columns)  $y_j$  from  $B_t$  (medoid approach), they construct a synthetic one, i.e., a virtual column  $c \in \mathfrak{R}^n$  in  $X$ . More specifically (and for centered columns  $y_j$ ), they define the prototype vector  $c_{B_t} := (c_{t1}, \dots, c_{tn})^\top \in \mathfrak{R}^n$  to be the vector  $c \in \mathfrak{R}^n$  that is most ‘similar’ to the variables in  $B_t$  in the sense

$$S(c; B_t) := \sum_{j \in B_t} r^2(y_j, c) = (1/n) c^\top X_{B_t} X_{B_t}^\top c \rightarrow \max_{c \in \mathfrak{R}^n, \|c\|=1}, \quad (5)$$

where  $X_{B_t}$  is the data matrix  $X$  restricted to the variables (columns) of  $B_t$ . Classical eigenvalue theory shows that the solution  $c_{B_t}$  is given by the standardized eigenvector  $v_t$  that belongs to the largest eigenvalue  $\lambda_t$  of  $X_{B_t} X_{B_t}^\top$  (and also  $X_{B_t}^\top X_{B_t}$ ), i.e., by the first principal component in  $B_t$ . Finally, Vigneau and Qannari formulate the following criterion for clustering variables:

$$g_3(\mathcal{B}; X) := \sum_{t=1}^{\ell} \sum_{j \in B_t} r^2(y_j, c_{B_t}) \rightarrow \max_{\mathcal{B}} \quad (6)$$

that is equivalent to the two-parameter *correlation clustering criterion*

$$g_4(\mathcal{B}, \mathcal{C}; X) := \sum_{t=1}^{\ell} \sum_{j \in B_t} r^2(y_j, c_t)^2 \rightarrow \max_{\mathcal{B}, \mathcal{C}} \quad (7)$$

where maximization is also with respect to the choice of the system  $\mathcal{C} = \{c_1, \dots, c_\ell\}$  of  $\ell$  standardized class-specific prototype variables (vectors)  $c_1, \dots, c_\ell \in \mathfrak{R}^n$ .

From its definition as a two-parameter optimization problems it is evident that for the variable clustering problem (7) a (sub-)optimum  $\ell$ -partition  $\mathcal{B}$  of variables can be obtained by a *generalized k-means algorithm*:

- (1) Begin with an initial partition  $\mathcal{B} = (B_1, \dots, B_\ell)$  of  $\mathcal{M} = \{1, \dots, p\}$ .
- (2) Partially optimize the clustering criterion with respect to the class prototype system  $\mathcal{C}$  for the classes  $B_t$  according to (5), thus yielding the class-specific eigenvector solutions  $c_{B_t}$  (class-specific principal components).
- (3) Build a new  $\ell$ -partition  $\mathcal{B}$  of the variables by assigning each variable  $y_j$  to the ‘most similar’  $c_{B_t}$ , i.e., the one with the largest value of  $r^2(y_j, c_{B_t})$ .
- (4) Iterate (2) and (3) until convergence.

Defining the similarity of variables by a correlation coefficient involves implicitly the concept of a linear regression. In fact, the correlation clustering criterion (7) above can be obtained from a probabilistic clustering model in which any variable  $y_j = (x_{1j}, \dots, x_{nj})^\top$  of a class  $B_t$  is generated, up to a random normal error, from the same latent factor (prototype variable)  $c_t = (c_{t1}, \dots, c_{tn})^\top \in \mathfrak{R}^n$  by a linear regression. The corresponding regression-type variable clustering model is given by

$$X_{ij} = a_j + b_j c_{ti} + e_{ij} \quad \text{for } i = 1, \dots, n; j \in B_t \quad (8)$$

with variable-specific intercepts  $a_j$ , slopes  $b_j$ , and independent normal errors  $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ . Estimating the unknown  $a_j, b_j$ , the prototype system  $\mathcal{C} = (c_1, \dots, c_\ell)$  and the  $\ell$ -partition  $\mathcal{B}$  by maximizing the likelihood of  $X = (x_{ij})$  is equivalent to the optimization problem

$$g_5(\mathcal{B}, \mathcal{C}, a, b; X) := \sum_{t=1}^{\ell} \sum_{j \in B_t} \sum_{i=1}^n (x_{ij} - a_j - b_j c_{ti})^2 \rightarrow \min_{\mathcal{B}, \mathcal{C}, a, b} \quad (9)$$

Partially optimizing the inner sum of  $g_5$  with respect to  $a_j, b_j$  yields the classical regression estimates

$$\hat{b}_j := \frac{s_{y_j c_t}}{s_{y_j y_j}} \quad \text{and} \quad \hat{a}_j = \bar{x}_{.,j} - \hat{b}_j \bar{c}_{t.} \quad \text{for } j \in B_t \quad (10)$$

in  $B_t$  with, e.g.,  $\bar{c}_{t.} := \sum_{i=1}^n c_{ti}/n$ , and the partial minimum of the two inner sums of (9) is given by

$$h(B_t, c_t) := \sum_{j \in B_t} \sum_{i=1}^n (x_{ij} - \hat{a}_j - \hat{b}_j c_{ti})^2 = \sum_{j \in B_t} n \cdot s_{y_j y_j} (1 - r^2(y_j, c_t))$$

and characterizes the homogeneity of  $B_t$  for a given prototype variable  $c_t$ . Finally, the multiparameter clustering problem (9) reduces to the two-parameter mixed continuous-discrete optimization problem for  $(\mathcal{B}, \mathcal{C})$ :

$$\begin{aligned} g_6(\mathcal{B}, \mathcal{C}; X) &:= \min_{a,b} g_5(\mathcal{B}, \mathcal{C}, a, b; X) = \sum_{t=1}^{\ell} h(B_t, c_t) \\ &= \sum_{t=1}^{\ell} \sum_{j \in B_t} n \cdot s_{y_j y_j} (1 - r^2(y_j, c_t)) \rightarrow \min_{\mathcal{B}, \mathcal{C}}. \end{aligned} \quad (11)$$

For the special case of standardized column variables  $y_j$ , i.e., for  $x_{.,j} = 0$  and  $s_{y_j y_j} = \|y_j\|^2/n = 1$ , this criterion is equivalent to the criterion (6) proposed by Vigneau and Qannari [43]. Insofar we have shown that their criterion (6) can be derived from a probabilistic clustering model. A software program in R is given by Chavent, Liquet, Kuentz-Simonet, and Saracco [15]. In the next section, a similar model will be used for modeling the co-clustering problem.

## 4 Co-Clustering, Where Variable Clusters are Characterized by Class-Specific Factors

In this section, we propose a co-clustering model for an  $n \times p$  object  $\times$  variable data table  $X = (x_{ij})$  with normally distributed entries where the clusters of objects (rows) are distinguished only by their levels (main effects) while each cluster of variables (columns) is, additionally, characterized by a cluster-specific factor with a high correlation to the variables within this class. Thereby we adopt the basic idea that has been followed in Sect. 3 when clustering the variables only. More specifically, with the notation of former sections and as an extension of the one-way clustering model (8), we consider the model

$$\begin{aligned} X_{ij} = \mu + \alpha_s + a_j + b_j c_{ti} + e_{ij} \quad & \text{for } i \in A_s, j \in B_t, \\ & s = 1, \dots, k, t = 1, \dots, \ell, \end{aligned} \quad (12)$$

where  $\mathcal{A} = (A_1, \dots, A_k)$  and  $\mathcal{B} = (B_1, \dots, B_\ell)$  are the unknown partitions of rows and columns, respectively (with known  $k$  and  $\ell$ ),  $\mu$  is a general effect and  $\alpha_s$  the ‘main effect’ of row class  $A_s$ . In this model, the vector  $c_t = (c_{t1}, \dots, c_{tn})^\top$  represents a cluster-specific latent factor that acts, in cluster  $B_t$ , as an explicative variable in the regression model that explains the  $n$  observations of variable  $j$  in the  $j$ -th column  $y_j = (x_{1j}, \dots, x_{nj})^\top$  of  $X$ , up to the main effects, by a linear regression  $a_j + b_j c_{ti}$



on the components of  $c_t$  with unknown variable-specific coefficients  $a_j$  and  $b_j$ . As before,  $e_{ij}$  are independent  $\mathcal{N}(0, \sigma^2)$  errors.

The clustering problem then consists in finding estimates for the parameters  $\mu$ ,  $\alpha = (\alpha_1, \dots, \alpha_k)$ ,  $a = (a_1, \dots, a_p)$ ,  $b = (b_1, \dots, b_p)$ ,  $\sigma^2$ , the set of factors  $\mathcal{C} = (c_1, \dots, c_\ell)$ , and the partitions  $\mathcal{A}$  and  $\mathcal{B}$  (under suitable norming constraints). In the model (12), the intercepts  $a_j$  of the linear regression part are specified separately for the variables  $j$ . In the following, we consider the more specialized co-clustering model where these intercepts are the same,  $\beta_t$  say, for all variables  $j$  from the same class  $B_t$ . This is described by the more specific co-clustering model

$$X_{ij} = \mu + \alpha_s + \beta_t + b_j c_{ti} + e_{ij} \quad \text{for } i \in A_s, j \in B_t, \quad (13)$$

$$s = 1, \dots, k, t = 1, \dots, \ell$$

with the constraints

$$\tilde{\alpha} := \sum_{s=1}^k \frac{|A_s|}{n} \alpha_s = 0, \quad \tilde{\beta} := \sum_{t=1}^{\ell} \frac{|B_t|}{p} \beta_t = 0, \quad \|c_t\|^2 = 1 \quad (14)$$

It describes a situation with additive class-specific main effects  $\alpha_s$  and  $\beta_t$  while interactions are cell-specific with the product form  $b_j c_{ti}$  (factor model).

Invoking the maximum likelihood approach for estimating the parameters in (13), we obtain the following factor-induced co-clustering problem:

$$Q(\mathcal{A}, \mathcal{B}; \mu, \alpha, \beta, b, \mathcal{C}; X)$$

$$:= \sum_{s=1}^k \sum_{t=1}^{\ell} \sum_{i \in A_s} \sum_{j \in B_t} (x_{ij} - \mu - \alpha_s - \beta_t - b_j c_{ti})^2 \rightarrow \min \quad (15)$$

where minimization is over all parameters under the constraints (14) (the model (12) may be treated similarly). In the following, we propose a *generalized alternating k-means-type algorithm* for solving this problem where, in each step, we partially optimize the criterion  $Q$ , (15), with respect to the involved parameters in turn.

*Step 1:* Choose an initial configuration  $(\mathcal{A}, \mathcal{B}, \mu, \alpha, \beta, \mathcal{C})$ . A reasonable choice might be  $\mu = \bar{x}_{\cdot, \cdot}$ ,  $\alpha_s = \bar{x}_{A_s, \cdot} - \bar{x}_{\cdot, \cdot}$ ,  $\beta_t = \bar{x}_{\cdot, B_t} - \bar{x}_{\cdot, \cdot}$ , while  $\mathcal{A}$  and  $\mathcal{B}$  could be obtained by separately clustering the rows and columns of  $X$ , e.g., by the classical  $k$ -means algorithm. Moreover, the class-specific factors  $c_1, \dots, c_\ell$  might be chosen randomly from the unit sphere in  $\mathfrak{R}^n$ .

*Step 2:* For fixed  $(\mathcal{A}, \mathcal{B}, \mu, \alpha, \beta, \mathcal{C})$ , determine the optimum regression coefficients  $b_1, \dots, b_\ell$  that minimize the criterion  $Q$ , (15). For notational convenience, we introduce the ‘adjusted’  $n \times p$  data matrix  $Z = (z_{ij}(\mathcal{A}, \mathcal{B}))$  with entries

$$z_{ij}(\mathcal{A}, \mathcal{B}) := z_{ij}(\mathcal{A}, \mathcal{B}; \mu, \alpha, \beta) := x_{ij} - \mu - \alpha_s - \beta_t$$

for  $i \in A_s, j \in B_t, s = 1, \dots, k, t = 1, \dots, \ell$

(where main effects are eliminated) such that this partial optimization problem takes the form

$$Q = \sum_{t=1}^{\ell} \sum_{j \in B_t} \sum_{i=1}^n (z_{ij}(\mathcal{A}, \mathcal{B}) - b_j c_{ti})^2 = \sum_{t=1}^{\ell} \sum_{j \in B_t} Q_j \rightarrow \min_{b_1, \dots, b_{\ell}}. \quad (16)$$

Minimizing, separately for each  $j \in B_t$ , the inner sum  $Q_j$  yields the estimates:

$$\hat{b}_j = \frac{\sum_{i=1}^n z_{ij}(\mathcal{A}, \mathcal{B}) c_{ti}}{\sum_{i=1}^n c_{ti}^2} = z_j(\mathcal{A}, \mathcal{B})^{\top} c_t \quad j \in B_t, t = 1, \dots, \ell$$

(with  $z_j(\mathcal{A}, \mathcal{B})$  the  $j$ -th column of  $Z$ ; note that  $\|c_t\|^2 = 1$ ) and the partial minimum

$$\begin{aligned} \tilde{Q}(\mathcal{C}) &:= \min_{b_1, \dots, b_{\ell}} Q = \sum_{t=1}^{\ell} \sum_{j \in B_t} \sum_{i=1}^n (z_{ij}(\mathcal{A}, \mathcal{B}) - \hat{b}_j c_{ti})^2 \\ &= \sum_{t=1}^{\ell} \sum_{j \in B_t} (\|z_j(\mathcal{A}, \mathcal{B})\|^2 - (z_j(\mathcal{A}, \mathcal{B})^{\top} c_t)^2) \end{aligned} \quad (17)$$

*Step 3:* Looking now for the factors  $c_t$  we have to minimize the criterion (17) with respect to  $\mathcal{C} = (c_1, \dots, c_{\ell})$ . This amounts to maximize, separately for each class  $B_t$ , the criterion

$$\sum_{j \in B_t} (z_j(\mathcal{A}, \mathcal{B})^{\top} c_t)^2 = c_t^{\top} \underbrace{\left[ \sum_{j \in B_t} z_j(\mathcal{A}, \mathcal{B}) z_j(\mathcal{A}, \mathcal{B})^{\top} \right]}_{S_t} c_t =: c_t^{\top} S_t c_t$$

with respect to  $c_t$  under the constraint  $\|c_t\| = 1$ . As in Sect. 3 the solution of this problem is given by the normalized eigenvector  $\hat{c}_t$  of the  $n \times n$  matrix  $S_t = S_t(\mathcal{A}, \mathcal{B}, \mu, \alpha, \beta)$  that belongs to the largest eigenvector of  $S_t$  (first principal component in  $B_t$ ).

*Step 4:* After having obtained the coefficients  $b_j = \hat{b}_j$  and the factors  $c_t = \hat{c}_t$  we substitute these estimates in the original co-clustering criterion  $Q$ , (15), and minimize it with respect to the global and main effects  $\mu, \alpha$ , and  $\beta$  under the norming constraints (14). A brief calculation yields the estimates:

$$\begin{aligned}\widehat{\mu} &= \bar{x}_{\cdot,\cdot} - \sum_{t=1}^{\ell} \frac{|B_t|}{p} \bar{b}_{B_t} \bar{c}_{t,\cdot} \quad \text{with } \bar{b}_{B_t} := \sum_{j \in B_t} \frac{b_j}{|B_t|} \\ \widehat{\alpha} &= \bar{x}_{A_s,\cdot} - \bar{x}_{\cdot,\cdot} - \sum_{t=1}^{\ell} \frac{|B_t|}{p} (\bar{c}_{t,A_s} - \bar{c}_{t,\cdot}) \\ \widehat{\beta}_t &= \bar{x}_{\cdot,B_t} - \bar{x}_{\cdot,\cdot} - \bar{b}_{B_t} \bar{c}_{t,\cdot} + \sum_{\tau=1}^{\ell} \frac{|B_\tau|}{p} \bar{b}_{B_\tau} \bar{c}_{t,\cdot}.\end{aligned}$$

While in Steps 2.–4. we have obtained the estimates for the effects  $\mu$ ,  $\alpha$ ,  $\beta$ , the coefficients  $b_j$  and the factors  $c_t$ , i.e., the configuration  $(\mathcal{A}, \mathcal{B}; \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{b}, \widehat{C})$  for a fixed choice of the partitions  $\mathcal{A} = (A_1, \dots, A_k)$  of objects and  $\mathcal{B} = (B_1, \dots, B_\ell)$  of variables, we now update these partitions by consecutively minimizing the criterion  $Q$ , (15), with respect to  $\mathcal{B}$  (Step 5.) and  $\mathcal{A}$  (Step 6.).

*Step 5:* Concerning first the partition  $\mathcal{B}$  of variables, the new and partially optimum  $\ell$ -partition  $\widehat{\mathcal{B}} = (\widehat{B}_1, \dots, \widehat{B}_\ell)$  for  $Q$  is the *minimum-distance partition* of  $\mathcal{M} = \{1, \dots, p\}$  with the classes

$$\widehat{B}_t := \{j \in \mathcal{M} \mid t = \operatorname{argmin}_{\tau=1, \dots, \ell} \delta(j, \tau; \mathcal{A}, \mathcal{B}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{b}, \widehat{C})\} \quad (18)$$

for  $t = 1, \dots, \ell$  where the distance measure  $\delta$  is defined by

$$\delta(j, \tau; \mathcal{A}, \mathcal{B}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{b}, \widehat{C}) := \|(z_j(\mathcal{A}, \mathcal{B}))\|^2 - (z_j(\mathcal{A}, \mathcal{B}))^\top c_\tau)^2 \quad (19)$$

for  $j = 1, \dots, p$ ,  $\tau = 1, \dots, \ell$  with  $z_{ij}(\mathcal{A}, \mathcal{B}) = z_{ij}(\mathcal{A}, \mathcal{B}; \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{b}, \widehat{C})$ . In fact, a look at (17) shows that the best partition  $\widehat{\mathcal{B}}$  has to minimize the distance  $\delta$ , (19), with respect to  $\tau$  for all variables  $j$ . Note that it follows from the original formula (15) for  $Q$  that the same partition is obtained when using the expression

$$\widetilde{\delta}(j, \tau; \mathcal{A}, \mathcal{B}, \widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{b}, \widehat{C}) := \sum_{s=1}^k \sum_{i \in A_s} (x_{ij} - \widehat{\mu} - \widehat{\alpha}_s - \widehat{\beta}_\tau - \widehat{b}_j \widehat{c}_{\tau i})^2$$

for  $j = 1, \dots, p$ ,  $\tau = 1, \dots, \ell$  instead of  $\delta$  in (18).

*Step 6:* Starting with the partition pair  $\mathcal{A}, \widehat{\mathcal{B}}$  and the current parameters  $\widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{b}, \widehat{C}$ , the estimation Steps 2.–4. are now repeated and will result in new estimates  $\mu^*, \alpha^*, \beta^*, b^*, C^*$ . With these estimates and the partition  $\widehat{\mathcal{B}}$  of variables, the  $k$ -partition  $\mathcal{A}$  of the set of objects  $\mathcal{O}$  is updated next: the new  $k$ -partition  $\widehat{\mathcal{A}}$  that partially minimizes the criterion  $Q(\mathcal{A}, \widehat{\mathcal{B}}; \mu^*, \alpha^*, \beta^*, b^*, C^*; X)$ , is the *minimum-distance partition* with classes

$$\widehat{A}_s := \{i \in \mathcal{O} \mid s = \operatorname{argmin}_{\sigma=1, \dots, k} d(i, \sigma; \mathcal{A}, \widehat{\mathcal{B}}, \mu^*, \alpha^*, \beta^*, b^*, C^*)\} \quad (20)$$

for  $s = 1, \dots, k$ , where the distance measure  $d$  is defined by

$$d(i, \sigma; \mathcal{A}, \hat{\mathcal{B}}, \mu^*, \alpha^*, \beta^*, b^*, C^*) := \sum_{t=1}^{\ell} \sum_{j \in B_t} (x_{ij} - \mu^* - \alpha_{\sigma}^* - \beta_t^* - b_j^* c_{ti}^*)^2 \quad (21)$$

for  $i = 1, \dots, n, \sigma = 1, \dots, k$ .

*Step 7:* The Steps 2.–6. are repeated until convergence of the two partitions.

Finally, we have obtained the partitions  $\mathcal{A}$  and  $\mathcal{B}$  of objects and variables (rows and columns), together with their characterizations, i.e.,

- the main effects  $\alpha_s$  of the classes  $A_s$  of objects;
- the main effects  $\beta_t$  of the classes  $B_t$  of variables together with the factors (prototype variables)  $c_1, \dots, c_{\ell} \in \mathfrak{R}^n$  of these classes.

The components of each factor  $c_t$  describe the contribution of the  $n$  objects to the composition of the column clusters  $B_t$  and the object  $\times$  variable interaction terms  $b_j c_{ti}$ . For easily interpreting the numerical results we can, e.g.,

- display, for each variable  $j$  from class  $B_t$ , the  $n$  points  $(c_{ti}, y_{ij})$ ,  $i = 1, \dots, n$ , in  $\mathfrak{R}^2$  that should be close to the corresponding regression line  $\eta = \beta_t + b_j c$ ;
- display and compare the latent factors  $c_1, \dots, c_{\ell}$  with the discrete curves  $(i, c_{ti})$ ,  $i = 1, \dots, n$ , in  $\mathfrak{R}^2$ , where the object labels  $i$  are arranged such object classes form contiguous segments; and
- visualize the  $\ell$  factors  $c_1, \dots, c_{\ell} \in \mathfrak{R}^n$  in a two-dimensional principal component display.

## 5 Discussion and Extensions

In this paper, we have proposed two probabilistic approaches for clustering simultaneously the objects (rows) and the variables (columns) of a data matrix. In contrast to other approaches where, e.g., ANOVA models or information distances are considered (see, e.g., Bock [8, 10–12]), our approach considers situations where the characterization of object clusters is different from the characterization of clusters of variables. In Sect. 2 this has been illustrated for the case when object clusters are characterized by class-specific means while variable clusters are characterized by class-specific variances. Moreover, in Sect. 4 we have introduced co-clustering models where object clusters were defined by main effects, and variable clusters by their main effects *and* a class-specific factor that explains the variables via a class-specific regression. This latter model was suggested after analyzing, in Sect. 3, a clustering method for variables only (proposed by Vigneau & Qannari [43]) and formulating a corresponding probabilistic model from which our new model can be derived.

For both co-clustering models, we have proposed an appropriate generalized  $k$ -means algorithm that proceeds by successively updating model parameters and partitions. These methods can be modified into various ways, e.g., by discussing the initial settings and the order of partial optimization steps. In this respect, this paper does not provide final results and lends itself to various investigations in the future. Basically, our models should be seen as a prototype for approaches that combine clustering of objects and clustering of variables in a simultaneous, probability-based framework. They can be extended to other two-parameter distributions, to the case of factor hyperplanes (involving higher principal components in each column class) and also to co-clustering models for three-way data similarly as in Bocci, Vicari, and Vichi [6], Schepers, Van Mechelen, and Ceulemans [37], Vichi, Rocci, and Kiers [42], or Wilderjans and Cariou [44], Wilderjans and Cariou [13].

## References

1. Aggarwal, C. C., & Reddy, C. K. (2014). *Data clustering. Algorithms and applications*. Boca Raton, Florida: CRC Press, Taylor & Francis.
2. Alfò, M., Martella, F., & Vichi, M. (2008). Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The International Journal of Biostatistics*, 4(1), Article 3.
3. Arabie, P., Schleutermann, S., Daws, J., & Hubert, L. (1988). Marketing applications of sequencing and partitioning on nonsymmetric and/or two-mode matrices. In W. Gaul & M. Schader (Eds.), *Data, expert knowledge and decisions* (pp. 215–224). Heidelberg: Springer Verlag.
4. Baier, D., Gaul, W., & Schader, M. (1997). Two-mode overlapping clustering with applications to simultaneous benefit segmentation and market structuring. In R. Klar, & O. Opitz (Eds.), *Classification and knowledge organization. Studies in Classification, Data Analysis, and Knowledge Organization* (vol. 9, pp. 557–566). Berlin, Germany: Springer.
5. Basu, S., Davidson, I., & Wagstaff, K. L. (2009). *Constrained clustering*. Boca Raton, Florida: Chapman & Hall/CRC, Francis & Taylor.
6. Bocci, L., Vicari, D., & Vichi, M. (2006). A mixture model for the classification of three-way proximity data. *Computational Statistics and Data Analysis*, 50, 1625–1654.
7. Bock, H.-H. (1974). *Automatische Klassifikation*. Göttingen: Vandenhoeck & Ruprecht.
8. Bock, H.-H. (1980). Simultaneous clustering of objects and variables. In R. Tomassone, M. Amirchahy, & D. Néel (Eds.), *Analyse de données et informatique* (pp. 187–203). Le Chesnay, France: INRIA.
9. Bock, H.-H. (1996). Probability models and hypothesis testing in partitioning cluster analysis. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 377–453). Singapore: World Scientific.
10. Bock, H.-H. (2003). Two-way clustering for contingency tables: Maximizing a dependence measure. In M. Schader, W. Gaul, M. Vichi (Eds.), *Between data science and applied data analysis. Studies in Classification, Data Analysis, and Knowledge Organization* (vol. 24, pp. 143–154). Berlin, Germany: Springer.
11. Bock, H.-H. (2004). Convexity-based clustering criteria: Theory, algorithms, and applications in statistics. *Statistical Methods & Applications*, 12, 293–314.
12. Bock, H.-H. (2016). Probabilistic two-way clustering approaches with emphasis on the maximum interaction criterion. *Archives of Data Science, Series A*, 1(1), 3–20.
13. Cariou, V., & Wilderjans, T. (2019). Constrained three-way clustering around latent variables approach. *Paper presented at the 16th conference of the International Federation of Classification Societies (IFCS-2019)*, Thessaloniki, Greece, 28 August 2019.0

14. Charrad, M., & Ben Ahmed, M. (2011). Simultaneous clustering: A survey. In S. O. Kuznetsov, et al. (Eds.), *Pattern recognition and data mining*, LNCS 6744 (pp. 370–375). Heidelberg: Springer Verlag.
15. Chavent, M., Liquet, B., Kuentz-Simonet, V., & Saracco, J. (2012). ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software*, 50, 1–16.
16. Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. In *Proceedings 8th international conference on intelligent systems for molecular biology* (pp. 93–103).
17. Cho, H., & Dhillon, I. S. (2008). Co-clustering of human cancer microarrays using minimum sum-squared residue co-clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3), 385–400.
18. Gaul, W., & Schader, M. (1996). A new algorithm for two-mode clustering. In H.-H. Bock & W. Polasek (Eds.), *Data analysis and information systems. Statistical and conceptual approaches. Studies in Classification, Data Analysis, and Knowledge Organization* (vol. 7, pp. 15–23). Heidelberg, Germany: Springer.
19. Dhillon, I. S. (2001). Co-clustering documents and words using bipartite graph partitioning. In *Proceedings of 7th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '01* (pp. 269–274). New York: ACM.
20. Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics*, 24(4), 437–458.
21. Govaert, G., & Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2), 463–473.
22. Govaert, G., & Nadif, M. (2008). Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6), 3233–3245.
23. Govaert, G., & Nadif, M. (2013). *Co-clustering*. Chichester, UK: Wiley.
24. Govaert, G., & Nadif, M. (2018). Mutual information, phi-squared and model-based co-clustering for contingency tables. *Advances in Data Analysis and Classification*, 12, 455–488.
25. Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2016). *Handbook of cluster analysis*. Boca Raton, Florida: CRC Press, Taylor & Francis.
26. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, New Jersey: Prentice Hall.
27. Kiers, H. A. L., Vicari, D., & Vichi, M. (2005). Simultaneous classification and multidimensional scaling with external information. *Psychometrika*, 70, 433–460.
28. Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE Transaction on Computational Biology and Bioinformatics*, 1(1), 24–45.
29. Martella, F., & Vichi, M. (2012). Clustering microarray data using model-based double k-means. *Journal of Applied Statistics*, 39(9), 1853–1869.
30. Martella, F., Alfò, M., & Vichi, M. (2010). Hierarchical mixture models for biclustering in microarray data. *Statistical Modelling*, 11(6), 489–505.
31. McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). Hoboken, New Jersey: Wiley.
32. Miyamoto, S., Ichihashi, H., & Honda, K. (2008). *Algorithms for fuzzy clustering*. Heidelberg: Springer Verlag.
33. Pontes, B., Giráldez, R., & Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. *ScienceDirect*, 57, 163–180.
34. Rocci, R., & Vichi, M. (2008). Two-mode partitioning. *Computational Statistics and Data Analysis*, 52, 1984–2003.
35. Salah, A., & Nadif, M. (2019). Directional co-clustering. *Advances in Data Analysis and Classification*, 13(3), 591–620.
36. Schepers, J., & Hofmans, J. (2009). TwoMP: A MATLAB graphical user interface for two-mode partitioning. *Behavioral Research Methods*, 41, 507–514.
37. Schepers, J., Van Mechelen, I., & Ceulemans, E. (2006). Three-mode partitioning. *Computational Statistics and Data Analysis*, 51, 1623–1642.
38. Schepers, J., Bock, H.-H., & Van Mechelen, I. (2013). Maximal interaction two-mode clustering. *Journal of Classification*, 34(1), 49–75.