

Advanced Information and Knowledge Processing

Editors-in-Chief

Lakhmi C. Jain, Bournemouth University, Poole, UK

University of South Australia, Adelaide, Australia

Xindong Wu, University of Vermont, USA

Series Editors

Sheryl Brahnam, Missouri State University, Springfield, USA

Diane J Cook, Washington State University, Pullman, WA, USA

Josep Domingo-Ferrer, Universitat Rovira i Virgili, Tarragona, Spain

Bogdan Gabrys, School of Design, Bournemouth University, Poole, UK

Francisco Herrera, ETS de Ingenierias Infoy de Telecom, University of Granada, Granada, Spain

Hiroshi Mamitsuka, School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan

Vir V. Phoha, Department of Electrical Engineering and Computer Science, Syracuse University, Ruston, LA, USA

Arno Siebes, Utrecht, The Netherlands

Philippe de Wilde, Office of the Vice Chancellor, University of Kent, Edinburgh, UK

Information systems and intelligent knowledge processing are playing an increasing role in business, science and technology. Recently, advanced information systems have evolved to facilitate the co-evolution of human and information networks within communities. These advanced information systems use various paradigms including artificial intelligence, knowledge management, and neural science as well as conventional information processing paradigms.

The aim of this series is to publish books on new designs and applications of advanced information and knowledge processing paradigms in areas including but not limited to aviation, business, security, education, engineering, health, management, and science.

Books in the series should have a strong focus on information processing - preferably combined with, or extended by, new results from adjacent sciences. Proposals for research monographs, reference books, coherently integrated multi-author edited books, and handbooks will be considered for the series and each proposal will be reviewed by the Series Editors, with additional reviews from the editorial board and independent reviewers where appropriate. Titles published within the Advanced Information and Knowledge Processing Series are included in Thomson Reuters' Book Citation Index and Scopus.

More information about this series at <https://link.springer.com/bookseries/4738>

Israël César Lerman · Henri Leredde

Seriation in Combinatorial and Statistical Data Analysis

Israël César Lerman
Data and Knowledge Management
Department
University of Rennes 1, IRISA
Rennes, Ille-et-Vilaine, France

Henri Leredde
Laboratoire Analyse Géométrie et
Applications (LAGA, CNRS UMR 7539)
Sorbonne Paris-Nord University
Villetaneuse, Seine-Saint-Denis, France

ISSN 1610-3947 ISSN 2197-8441 (electronic)
Advanced Information and Knowledge Processing
ISBN 978-3-030-92693-9 ISBN 978-3-030-92694-6 (eBook)
<https://doi.org/10.1007/978-3-030-92694-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed 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 Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

To begin, let us describe the circumstances and motivations that led to the publication of this book. This publication can be related to that of the book entiteled

*Foundations and Methods in Combinatorial and Statistical
Data Analysis and Clustering*

published in the same Springer Nature series in 2016. Let us recall that the latter book was a new recasting and comprehensive english version of an important part of the french book:

Classification et Analyse Ordinale des Données

published with the support of the CNRS—by Dunod (Paris) in 1981.

Clearly, in this version, the immense development of the domain has been taken into account. It was Dan A. Simovici, Professor at the University of Massachusetts (Department of Computer Science) who encouraged me to propose this publication. I am very grateful to him.

In the development of the English book above, denoting by E the set to be statistically structured, we are led to provide E with a *symmetrical* numerical or ordinal similarity or dissimilarity over E . This symmetry is with respect to $E \times E$. On the other hand and above all, the synthetic structure built on E is *symmetrical*. The latter structure is defined by either a classification tree (hierarchical clustering) or by a classification (non-hierarchical clustering) of E .

In combinatorial data analysis, the synthetic structure may be ordinal. In this case, relative to a given ordered pair (x, y) in $E \times E$, x is to y what is not y to x .

The initial idea of a second volume was to concentrate the matter on methods for which the synthetic structure built on E is *asymmetrical*. Mostly, this relation has an ordinal nature. It might be discrete (logical) or valuated, numerically. Now, regarding the initial similarity (resp., dissimilarity) measure given on E , the latter might be symmetrical or oriented. The subjects of ordinal data analysis which are taken into account in the french book (Dunod 1981) are:

- Principal Component Analysis and Correspondence Analysis;
- Formal and Methodological Comparison Between Component Factorial Approach and Classification Approach;
- Combinatorial and Statistical Seriation Methods in Relation to a Family of Cluster Analysis Methods;
- Totally ordering the whole set of categories associated with a set of ordinal categorical attributes;
- Assignment problems in *Pattern Recognition* between geometrical figures where the quality measure of the assignment has to be independent of specific geometrical transformations applied on the figures concerned.

In this context, an additional chapter may be defined by the description and analysis of *Directed Binary Hierarchy*. The latter corresponds to an asymmetrical structure in which the leaf set is sustained by a linear order. To situate this structure—introduced by Régis Gras—with respect to the classical and then symmetrical binary hierarchy, it was formalized and studied in the first chapter of the Springer Nature book volume mentioned above, and previously, in an important article “Directed Binary Hierarchies and Directed Ultrametrics” by Israël César Lerman and Pascale Kuntz in *The Journal of Classification* (28): 272–296 (October 2011).

Resume in one book all of the themes that we have just quoted above would have required a too large volume whose material may appear as too scattered. In this book, we prefer to focus on the subject of *Seriation*. This corresponds to two of the chapters of the French book mentioned above. It is indeed a rather broad subject in its own right. On the other hand, the seriation approach has close links with Clustering (Classification). Both approaches have to be situated facing each other. This is the reason of the title of this book.

The works just mentioned have already given rise—apart from what is included in the french book—to a thesis [1] and the article [2]. The book proposed here goes very far beyond these works:

- The mathematical results are more accurate and richer;
- New and powerful algorithmic combinatorial and statistical methods are proposed and validated;
- A very rich synthesis is proposed, the new methods are situated in relation to a very important set of recent methods;
- A vast experimentation on simulated or real data is carried out.

Let us now briefly describe the respective contents of the different chapters.

The first chapter introduces the problem of seriation, the methods proposed to treat it, as well as the history of the evolution of the domain.

In the second chapter, a formal mathematical expression is proposed in the case where the data is defined by an incidence table composed of zeros and ones and therefore, where the descriptive attributes are Boolean. In this chapter, we establish the statistical bases leading to the development of an original and very simple methodology for planar geometric representation of the columns or rows of the incidence

table. This method, called *Attraction Pole* method enables a seriation to be revealed. Its behavior will be experimented in Chap. 4.

Chapter 3 offers a detailed and commented description of the main methods considered in the literature to attack the seriation problem. The mutual links between the different approaches are brought out. This chapter ends by indicating a combinatorial version for attraction pole method. The latter will be developed in Chap. 5.

As just mentioned, Chap. 4 develops an experimental analysis comparing several methods of planar geometric representation: the new one of the poles of attraction and the classic and proven ones. For the latter, it is the principal component analysis, correspondence analysis and multidimensional scaling that underlies D. Kendall's Horse-shoe method (see [3]). Several seriation forms are tested, some of them include two or even three blocks.

In Chap. 5, a new family of ordinal and combinatorial seriation algorithms is established. In these, the seriation result depends on how to choose to start the order. The behavior of these algorithms is experimentally proven with respect to simulated or real data.

The determination of a system of attraction poles, mutually distant from each other enables a new and rich family of clustering methods to be derived. This is developed in Chap. 6. The respective relationships of this new family with that of the *K-means* on the one hand, and that of ascendant hierarchical classification on the other, are established.

In conclusion, we will give the possible extensions of the whole work.

Rennes, France
Villetaneuse, France

Israël César Lerman
Henri Leredde

References

1. H. Leredde, La méthode des pôles d'attraction, La méthode des pôles d'agrégation. PhD thesis, Université de Paris 6, October 1979
2. I. C. Lerman, H. Leredde, La méthode des pôles d'attraction, in *Analyse des Données et Informatique*, ed. by E. Diday et al. (IRIA, 1977), pp. 37–49
3. D. G. Kendall, Seriation from abundance matrices, in *Mathematics in Archaeological and Historical Sciences*, ed. by D. G. Kendall, F. R. Hodson, P. Tautu (Aldine-Atherton, Chicago, 1971), pp. 214–252

Acknowledgements

We acknowledge with gratitude the support of various institutions and individuals who have sponsored our work. The University of Rennes 1 and the University of Sorbonne Paris Nord have provided support in the framework of their respective departments the ISTIC (Informatique et Communications) and the Galilée Institute with Mathematics Department and his laboratory LAGA.

Gilles Lesventes (University of Rennes 1) and Jean-Pierre Astruc (President of the Sorbonne Paris Nord University) welcomed and supported us with great kindness.

Extensive facilities have been provided by the IRISA (Institut de Recherche en Informatique et Systèmes Aléatoire) Institute. We especially thank their directors Bruno Arnaldi, Jean Marc Jezequel and Guillaume Gravier.

Philippe Louarn (INRIA-Rennes) has defined the general LaTeX structure with respect to which we have composed this book. He helped us many times and his help was always valuable. We are very grateful to him.

We also have immense gratitude to the editor of the series in which this book is published as well to the editorial staff of Springer Nature, particularly to Helen Desmond who has always been in our listening.

Contents

1	General Introduction: Methods and History	1
	References	8
2	Seriation from Proximity Variance Analysis	11
2.1	Introduction	11
2.2	Definition of a Seriation; Its Unicity	13
2.2.1	Definition of a σ Form	13
2.2.2	Unicity	15
2.3	Association Coefficient Between Columns of a σ Seriation	
	Form	17
2.3.1	Preamble	17
2.3.2	The Association Coefficient	18
2.3.3	Equation of the Association Coefficient in the Case of a σ Form	20
2.3.4	Representation on a Directed Unit Line Segment	24
2.4	Simultaneous Mean and Variance Analysis of Pairwise Proximities	25
2.4.1	Preamble	25
2.4.2	Optimal Properties of an Extreme Point of Rectilinear Cloud	26
2.4.3	Discovering from an Optimal Property, a σ Form	32
2.4.4	Block Seriation from Simultaneous Mean and Variance Analysis of Pairwise Proximities; The “Attraction Poles” Method	39
2.4.5	Proximity Variance Analysis Equation	43
2.4.6	Geometrical Representation by the Attraction Poles Method	47
	References	56

3	Main Approaches in Seriation: The Attraction Pole Case	57
3.1	Introduction	57
3.2	Visual and Combinatorial Methods	61
3.2.1	Graphical Methods of J. Bertin	61
3.2.2	Block Seriation Method of F. Marchotorchino	63
3.2.3	Methodological Developments	64
3.2.4	Elisséeff Heuristic	66
3.2.5	The Deutsch and Martin Algorithm	72
3.2.6	The Niermann Algorithm	74
3.3	Seriation Defined as a Combinatorial Optimization Problem	78
3.3.1	Preamble	78
3.3.2	General Facets	78
3.3.3	A Criterion for Evaluating the Seriation	85
3.3.4	Usual Criteria for Evaluating a Seriation	90
3.4	Spectral Approaches	96
3.4.1	Some Methodological Points and Extensions	99
3.5	Methods Using a Planar Geometrical Representation	102
3.5.1	Introduction: The Most Classical Methods	102
3.5.2	The Horse-Shoe Method of Kendall	103
3.6	Combinatorial and Algorithmic Approaches of the Pole Attraction Method	107
	References	108
4	Comparing Geometrical and Ordinal Seriation Methods in Formal and Real Cases	111
4.1	Introduction	111
4.2	Methods	113
4.2.1	Similarities and Distances	113
4.2.2	Multidimensional Data Analysis	115
4.3	Techniques and Algorithms	116
4.4	Results in Processing Data	119
4.4.1	Preamble	119
4.4.2	Simulated Data According to σ Forms Models	120
4.4.3	Real Data	147
4.4.4	Some Concluding Remarks	171
	References	173
5	A New Family of Combinatorial Algorithms in Seriation	175
5.1	Introduction: The Fundamental Principles	175
5.2	Row Data Table Ranking Associated with a Column Ordering in Seriation	177
5.2.1	Introduction	177
5.2.2	The Combinatorial Algorithm	179
5.2.3	The Statistical Algorithm	181

5.3	Seriation Algorithms	183
5.3.1	The Seriation Search Space Is Associated with the First Element	183
5.3.2	The Seriation Search Space Is Associated with Sized Sequences of Elements	186
5.4	Applying on σ Forms and Real Data	195
5.4.1	Preamble	195
5.4.2	Simulated σ Forms	196
5.4.3	Real Data	203
	References	213
6	Clustering Methods from Proximity Variance Analysis	215
6.1	Introduction and General Presentation	215
6.2	The Method Family of Attraction Poles by Successive Reallocations (<i>APMSR</i>)	217
6.2.1	Data, Similarities and Distances	217
6.2.2	Determination of a System of Attraction Poles	220
6.2.3	Cluster Formation Around the Attraction Poles	222
6.2.4	Assignment Criteria	223
6.3	Criteria and Algorithmic of the Attraction Pole Methods by Successive Aggregations (<i>APMSA</i>)	229
6.3.1	Introduction	229
6.3.2	Criteria for <i>APMSA</i>	231
6.4	On the Number of Poles to Extract and on the Quality of the Clusterings Obtained	232
6.4.1	The Fitting Criterion	233
6.4.2	Inertia or Ward Criterion	234
6.4.3	Interesting Partitions from <i>APMSR</i> and from <i>APMSA</i>	236
6.5	Developments	237
6.6	Applying Clustering Attraction Pole Methods in Real Data	239
6.6.1	Introduction	239
6.6.2	Ruspini Data	241
6.6.3	Fisher Data	250
6.6.4	Merovingian Belt Buckles-Plates Data	254
6.6.5	Some Words to Conclude	260
	References	261
7	Conclusion and Developments	263
7.1	Introduction	263
7.2	The Algorithms and Their Respective Logics	264
7.2.1	Geometric Representations of the Descriptive Attributes and the Objects Described	264
7.2.2	Row Seriation from Column Seriation	265
7.2.3	Algorithms by Successive Chaining	265
7.2.4	Attraction Pole Clustering Algorithms	266

7.3	Structures and Algorithms in Asymmetrical Hierarchical	
	Clustering	266
7.3.1	Directed Ascendant Hierarchical Clustering	266
7.3.2	Oriented Hierarchical Clustering	268
7.3.3	Hierarchical Clustering of Successive Intervals	269
7.3.4	Some Additional Interactions Between Seriation	
	and Hierarchical Clustering	270
References	271
Index	273