SPRINGER BRIEFS IN STATISTICS

Ingwer Borg
Patrick J. F. Groenen
Patrick Mair

Applied Multidimensional Scaling and Unfolding

Second Edition

EXTRAS ONLINE

Springer

SpringerBriefs in Statistics

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

SpringerBriefs in Statistics showcase emerging theory, empirical research, and practical application in Statistics from a global author community.

SpringerBriefs are characterized by fast, global electronic dissemination, standard publishing contracts, standardized manuscript preparation and formatting guidelines, and expedited production schedules.

More information about this series at http://www.springer.com/series/8921

Ingwer Borg · Patrick J. F. Groenen
Patrick Mair

Applied Multidimensional Scaling and Unfolding

Second Edition

Springer

Ingwer Borg
Westfälische Wilhelms-Universität
Münster
Germany

Patrick J. F. Groenen
Econometric Institute
Erasmus University Rotterdam
Rotterdam
The Netherlands

Patrick Mair
Department of Psychology
Harvard University
Cambridge, MA
USA

Additional material to this book can be downloaded from http://extras.springer.com.

ISSN 2191-544X ISSN 2191-5458 (electronic)
SpringerBriefs in Statistics
ISBN 978-3-319-73470-5 ISBN 978-3-319-73471-2 (eBook)
https://doi.org/10.1007/978-3-319-73471-2

Library of Congress Control Number: 2018934861

Mathematics Subject Classification (2010): 91C15

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Multidimensional scaling (MDS) is a powerful statistical method that maps proximity data on pairs of objects (i.e., data expressing the similarity or the dissimilarity of pairs of objects) into distances between points in a multidimensional space. The space is usually two-dimensional, sometimes also three-dimensional, and seldom more than three-dimensional. Unfolding is a related method for preference data (e.g., persons' ratings on choice objects such as consumer goods). It maps these data into distances between points representing the persons and points representing the choice objects.

The purpose of MDS and unfolding is often just visualizing the data so it becomes easier for the user to explore and to understand their structure. However, both MDS and unfolding can also be used to test a variety of structural hypotheses about the data or even psychological theories of judgment or choice. Thousands of publications have used MDS and unfolding in these ways.

This book is a brief introduction to MDS and unfolding. It discusses the issues that always come up when MDS or unfolding is used in substantive research, and it shows how to actually run such analyses. The aim is conceptual understanding and practical know-how rather than mathematical precision and proof. It is more like a driving lesson, not like engineering a car. These are different things, and the engineer is not necessarily a better driver.

In this second edition, we focus much more on R packages and the R environment than we did in the first edition. However, we decided not to drop other computer packages (such as SPSS and its modules, in particular), because many users are (still?) using these programs. Moreover, some of these programs have features that are not available in R yet. On the other hand, we mention highly special stand-alone programs only occasionally, since many of them are hard to get and difficult to use.

This edition also puts much more emphasis on unfolding. Unfolding was almost completely neglected in the first edition, since nobody used it, even though it is a powerful method and an interesting model. Things have changed recently: Unfolding seems to become more popular in substantive research and in consulting.

With regard to MDS, we introduce and explain recent developments that are concerned with the goodness of an MDS solution and with its substantive interpretation. They are particularly important for the MDS user, for reviewers, and for journal editors. For example, MDS users can now test the statistical significance of MDS (and unfolding) solutions using methods that require computer simulations that were difficult to run within traditional statistics packages but that are now easily feasible within the R environment.

We also present various new examples of how to run an MDS or an unfolding job using R. These examples are almost all substantively relevant and not just contrived illustrative examples. Most data that we use in this book are also readily available in the R package SMACOF so that the user can check our analyses.

To make our cases as concrete as possible, we repeatedly show R scripts for running the jobs. In these scripts, we tried adhering to the R etiquette of writing R code, but did not follow it strictly where it would waste too much space. For example, we often use the semicolon to write more than just one command per line. Prettier code can easily be generated by marking the code and then typing Ctrl +Shift+A in RStudio, for example, or by using the `tidy_source()` function in the formatR package. The scripts shown in this book (and a few additional ones) are also available, in prettier form, in the supplementary script file. Additional material to this book can be downloaded from http://extras.springer.com. It should also be noted that some plots do not correspond exactly to those produced by the various scripts. Rather, some plots were slightly edited by hand to unclutter, in particular, the labels attached to the points in scatter plots.

Münster, Germany — Ingwer Borg
Rotterdam, The Netherlands — Patrick J. F. Groenen
Cambridge, USA — Patrick Mair

Contents

Chapter 1
First Steps

Abstract The basic ideas of MDS are introduced doing MDS by hand. Then, MDS is done using statistical software. The goodness of the MDS configuration is evaluated by correlating its distances with the data. Unfolding is introduced with a small example.

Keywords MDS · Iteration · Proximities · Dimensional interpretation
Goodness of fit · Unfolding

1.1 Basic Ideas of Multidimensional Scaling

The basic ideas of MDS are easily explained using a small example. Consider Table 1.1. It contains the correlations of different crimes in 50 US states. The correlations show, for example, that if there are many cases of Assault in a state, then there are also many cases of Murder ($r = 0.81$). We now scale these correlations via MDS. This means that we try to represent the seven crimes by seven points in a geometric space so that any two points lie the *closer* together the *greater* the correlation of the crimes that these points represent.

To reach this goal, we take seven cards, and write the name of one crime on each of them, from Murder to Auto Theft. These cards are then placed on a table in an arbitrary arrangement (Fig. 1.1). Their distances are measured (Fig. 1.2) and compared with the correlations in Table 1.1. This comparison shows that the configuration in Fig. 1.1 does not represent the data in the desired sense. For example, the cards for Murder and Assault should be relatively close together, because these crimes are correlated with 0.81, whereas the cards for Murder and Larceny should be farther apart, as these crimes are correlated with only 0.06. We, therefore, try to move the cards repeatedly in small steps ("iteratively") so that the distances correspond more closely to the data. Figure 1.3 demonstrates in which directions the cards should be shifted to improve the correspondence of data and distances.

Improving a given configuration iteratively by hand can be fairly tedious. It also does not guarantee convergence to a stable and optimal configuration. So, let an

I. Borg et al., *Applied Multidimensional Scaling and Unfolding*,
SpringerBriefs in Statistics, https://doi.org/10.1007/978-3-319-73471-2_1

Table 1.1: Correlations of crime rates in 50 US states

Crime	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto Theft
Murder	1.00	0.52	0.34	0.81	0.28	0.06	0.11
Rape	0.52	1.00	0.55	0.70	0.68	0.60	0.44
Robbery	0.34	0.55	1.00	0.56	0.62	0.44	0.62
Assault	0.81	0.70	0.56	1.00	0.52	0.32	0.33
Burglary	0.28	0.68	0.62	0.52	1.00	0.80	0.70
Larceny	0.06	0.60	0.44	0.32	0.80	1.00	0.55
Auto Theft	0.11	0.44	0.62	0.33	0.70	0.55	1.00

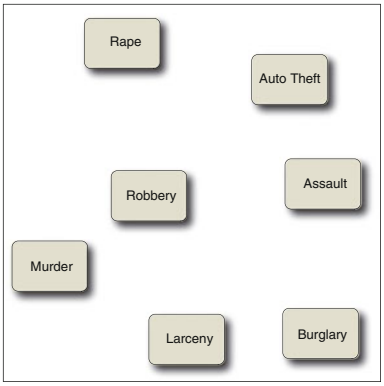

Fig. 1.1: Initial configuration for an MDS of the data in Table 1.1

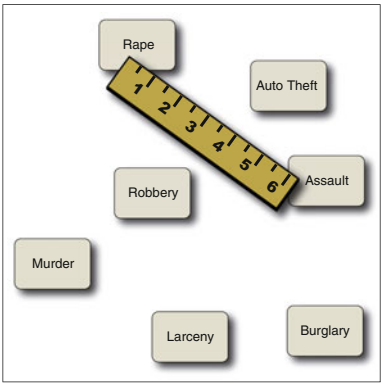

Fig. 1.2: Measuring distances with a ruler

MDS computer algorithm do the job. It systematically moves the points step by step to improve the fit to the data.

There exist many good MDS programs. One such program is PROXSCAL, a module of SPSS. To use PROXSCAL, we first save the correlation matrix of Table 1.1 in a file that we call 'CorrCrimes.sav'. Then, we only need some clicks in PROXSCAL's menus (click: Analyze > Scale > Multidimensional Scaling (PROXSCAL)) or, alternatively, execute the following commands:

```
GET FILE='CorrCrimes.sav'.
PROXSCAL VARIABLES=Murder to AutoTheft
     /TRANSFORMATION=INTERVAL
     /PROXIMITIES=SIMILARITIES .
```

The `PROXIMITIES` sub-command informs the program that the data—called *proximities* in this context, a generic term for both *similarity* and *dissimilarity* data—must be interpreted as similarities by the program. That is, small data values should be mapped

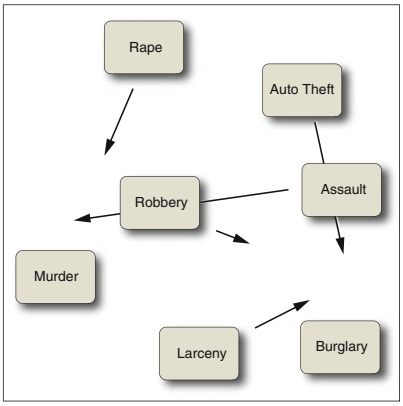

Fig. 1.3: Directions for point movements to improve the MDS configuration

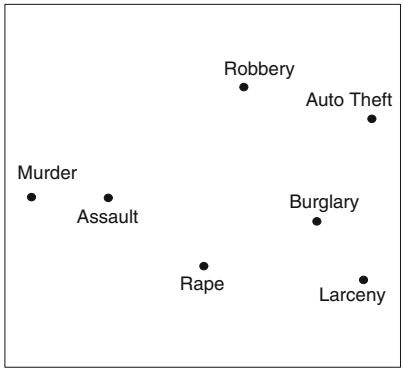

Fig. 1.4: Optimal computer-generated MDS solution

into large distances, and large data values into small distances. Also, we want to map the correlations linearly into MDS distances, preserving their differences ("intervals") in the distances. In PROXSCAL, we thus request `/TRANSFORMATION=INTERVAL`. No further specifications are needed. The program uses its default settings to generate an MDS solution (Fig. 1.4).

Many other programs exist for MDS. One example is the MDS module in SYSTAT. SYSTAT can be run using commands, or by clicking on various options in a graphical user interface. Having loaded the correlation matrix as our data, we call the MDS module and its menu in Fig. 1.5. We select the variables Murder, Rape, etc., and leave all other specifications as they are, except the one for "Regression", where we request that the MDS program should optimize the relation of data to distances in the sense of a least-squares *linear* regression. Clicking on the OK button makes the program find and plot an MDS configuration.

A third implementation is the `mds()` function of the R (R Core Team 2017) package SMACOF (De Leeuw and Mair 2009). SMACOF is open source and, most importantly, allows using the sheer boundless capabilities of the R environment and its thousands of software packages for additional analyses, simulations, and graphics. So, we will mostly use SMACOF in this book.

SMACOF is run by commands. A few commands suffice to do the MDS analysis of the given data. Note that SMACOF always requires that the data either come as dissimilarities, or that they have been converted to dissimilarities (accomplished here by the `sim2diss` function).