Gifi Analysis of Multivariate Data

Jan de Leeuw

First created May 04, 2016. Last update October 09, 2023

Contents

}

}

Note

This book will be expanded/updated frequently. The directory deleeuwpdx.net/pubfolders/stress has a pdf version, the bib file, the complete Rmd file with the code chunks, and the R and C source code. Suggestions for improvement of text and code are welcome. All text and code are in the public domain and can be copied and used by anybody in any way they like. Attribution will be appreciated, but is not required.

Just as an aside: "above" in the text refers to anything that comes earlier in the book and "below" refers to anything that comes later. This always confuses me, so I had to write it down. I also number *all* displayed equations. Equations are displayed if and only if they are important, are referred to in the text, or mess up the line spacing.

Preface

In 1980 members of the Department of Data Theory at the University of Leiden taught a post-doctoral course in Nonlinear Multivariate Analysis. The course content was sort-of-published, in Dutch, as Gifi (1980). The course was repeated in 1981, and this time the sort-of-published version (Gifi (1981)) was in English.

The preface gives some details about the author.

> The text is the joint product of the members of the Department of Data Theory of the Faculty of Social Sciences, University of Leiden. 'Albert Gifi' is their 'nom de plume'. The portrait, however, of Albert Gifi shown here, is that of the real Albert Gifi to whose memory this book is dedicated, as a far too late recompense for his loyalty and devotion, during so any years, to the Cause he served.

Roughly ten years later a revised version of these course notes came out as an actual book in the *Wiley Series in Probabilty and Mathematical Statistics* (Gifi (1990)). This despite the fact that the contents of the book had very little to do with either probability or mathematical statistics. The book is organized around a series of computer programs for correspondence analysis, principal component analysis, and canonical analysis. The programs, written in FORTRAN, are called HOMALS, PRINCALS, PRIMALS, CRIMINALS, CANALS, OVERALS because they combine classical linear multivariate analysis with optimal transformation of the variables, using alternating least squares (or ALS). It serves, to some extent, as a manual for the programs, but it also discusses the properties of the techniques implemented in the programs, and it presents many detailed applications of these techniques.

Reviewers generally had some difficulties separating the wheat from the chaff.

> As the spirit of Albert Gifi has faded away, so has his whimsical approach to publishing, and his latest book is an idiosyncratic account of multivariate methods developed by the Leiden group during the 1970s. The names of their computer programs are distinguished by the ending ~ALS, thus we have OVERALS, PRINCALS, HOMALS, CANALS, MORALS, MANOVALS, CRIMINALS, PARTALS and PATHALS. Perhaps if you have a warped mind like this reviewer, you will turn rapidly to CRIMINALS. What can it be ? Surely it must give some illicit view of the truth about the world, a vision of the underworld of multivariate analysis ? Alas no ! It turns out only to be a synonym of Canonical Variate Analysis, sometimes known as Multiple Discriminant Analysis. Likewise HOMALS turns out to be Reciprocal Averaging, otherwise known as Correspondence Analysis. (Hill (1990))

This ambiguity and confusion are not too surprising. The Gifi book was a summary of the work of a large number of people, over a period of almost 20 years. Nevertheless, and perhaps because of this, it is somewhat of a *camel*, which we define for our purposes as a *horse designed by a committee*. Different chapters had different authors, and the common ideas behind the various techniques were not always clearly explained.

> In Gifi's MVA the criterion called "meet" loss plays a central role. Although the adoption of this criterion is one of the most important contributions of Gifi, the book would have been much more readable if this criterion had been introduced right at the outset and was followed throughout the rest of the book. (Takane (1992))

Nevertheless there is much original material in Gifi (1990), and the book has early applications of alternating least squares, majorization, coordinate descent, the delta method, and the bootstrap. And it emphasizes throughout the idea that statistics is about techniques, not about models. But, yes, the organization leaves much to be desired. An on demand printing of the first and only edition is now available on Amazon for $ 492 – although of course used versions go for much less.

The book was published by a prestiguous publisher in a prestiguous series, but it is fair to say it never really caught on. It is not hard to understand why. The content, and the style, are unfamiliar to statisticians and mathematicians. There is no inference, no probability, and very little rigor. The content is in multivariate data analysis, which would be most at home these days, if anywhere, in a computer science department. The Gifi group did not have the resources of, say, Benzécri in France or Hayashi in Japan. The members were mostly active in psychometrics, a small and insular field, and they were from The Netherlands, a small country prone to overestimate its importance (Marvell (1653)). They also did not have the evangelical zeal necessary for creating and sustaining a large impact.

There have been some other major publication events in the Gifi saga. Around the same time as the Wiley book there was the publication of SPSS (1989). Starting in the late seventies the Gifi FORTRAN programs had been embedded in the SPSS system. The *SPSS Categories* manual was updated many times, in fact every time SPSS or IBM SPSS had a new release. Over the years other programs produced by the Department of Data Theory were added. A recent version is, for example, Meulman and Heiser (2012), corresponding to IBM SPSS 21. It acknowledges the contributions of some of the members of the Gifi team – but in IBM (2015), the version for IBM SPSS 23, these acknowledgements and the names of the authors have disappeared. Sic transit gloria mundi.

Michailidis and De Leeuw (1998) made an attempt to make the Gifi material somewhat more accessible by publishing a review article in a widely read mainstream statistical journal. Another such attempt is De Leeuw and Mair (2009), in which the homals package for R is introduced. The homals package is basically a single monolithic R function that can do everything the Gifi programs can do, and then some. In both cases, however, the problem remained that the techniques, and the software, were too convoluted and too different from what both statisticians and users were accustomed to.

Van der Heijden and Van Buuren (1916) give an excellent, though somewhat wistful, historic overview of the Gifi project. It is too early for eulogies, however, and we refuse to give up. This book is yet another reorganization of the Gifi material, with many extensions. We take Yoshio Takane's advice seriously, and we organize both the theory and the algorithms around what is called "meet-loss" in Gifi. In our software we separate the basic computational engine from its various applications that define the techniques of *Multivariate*

Analysis with Optimal Scaling (MVAOS). Hiding the core makes it possible to make the programs behave in much the same way as traditional MVA programs. The software is written in R ((**?**)), with some parts of the computational engine written in C.

The book itself is written in Rmarkdown, using bookdown (Xie (2016)) and knitr (Xie (2015)) to embed the computations and graphics, and to produce html and pdf versions that are completely reproducible. The book and all the files that go with it are in the public domain.

We would like to acknowledge those who have made substantial contributions to the Gifi project (and its immediate ancestors and offspring) over the years. Some of them are lost in the mists of time, some of them are no longer on this earth. They are, in alphabetical order, Bert Bettonvil, Jason Bond, Catrien Bijleveld, Frank Busing, Jacques Commandeur, Henny Coolen, Steef de Bie, Jan de Leeuw, John Gower, Patrick Groenen, Chris Haveman, Willem Heiser, Abby Israels, Judy Knip, Jan Koster, Pieter Kroonenberg, Patrick Mair, Adriaan Meester, Jacqueline Meulman, George Michailidis, Peter Neufeglise, Dré Nierop, Ineke Stoop, Yoshio Takane, Stef van Buuren, John van de Geer, Gerda van den Berg, Eeke van der Burg, Peter van der Heijden, Anita van der Kooij, Ivo van der Lans, Rien van der Leeden, Jan van Rijckevorsel, Renée Verdegaal, Peter Verboon, Susaññа Verdel, and Forrest Young.

Chapter 1

Introduction

Placeholder

1.1 Some Dualisms

1.2 Quantifying Qualitative Data

1.3 Beyond Gifi

Chapter 2

Coding and Transformations

Placeholder

2.1 Variables and Multivariables

2.2 Induced Correlations and Aspects

2.3 Transformed Variables

2.4 Bases

2.5 Copies and Rank

2.6 Orthoblocks

2.7 Constraints

2.8 Missing Data

2.9 Active and Passive Variables

2.10 Interactive Coding

Chapter 3

Aspects

Placeholder

3.1 Definition

3.2 Stationary Equations

3.3 Bilinearizability

Chapter 4

Pattern Constraints and Gifi Loss

Placeholder

4.1 Aspects from Patterns

4.2 Gifi Loss

4.3 Associated Eigenvalue Problems

4.4 History

Chapter 5

Algorithm

Placeholder

5.1 Block Relaxation

5.2 Majorization

5.3 Alternating Least Squares

5.4 Implementation Details

5.5 Wrappers

5.6 Structures

Chapter 6

Multiple Correspondence Analysis and homals()

Placeholder

6.1 Introduction

6.2 Equations

6.3 Examples

6.3.1 Hartigan's Hardware

6.3.2 Thirteen Personality Scales

Chapter 7

Canonical Correspondence Analysis and coranals()

7.1 Introduction

7.2 Equations

Canonical analysis of GX and H.

7.3 Examples

Chapter 8

Nonlinear Principal Component Analysis and princals()

Placeholder

8.1 Introduction

8.2 Equations

8.3 Examples

8.3.1 Thirteen Personality Scales

Chapter 9

Canonical Analysis and canals()

9.1 Equations

If there are only two blocks the generalized eigenvalue problem for the Burt matrix becomes

$$\begin{bmatrix} D_1 & C_{12} \\ C_{21} & D_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = 2\lambda \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix},$$

which we can rewrite as

$$\begin{aligned} C_{12}a_2 &= (2\lambda - 1)D_1a_1, \\ C_{21}a_1 &= (2\lambda - 1)D_2a_2, \end{aligned}$$

from which we see that MVAOS maximizes the sum of the r largest canonical correlations between H_1 and H_2. See also Van der Velden (2012).

9.2 Examples

Chapter 10

Multiple Regression and morals()

Placeholder

10.1 Equations

10.2 Examples

10.2.1 Polynomial Regression

10.2.2 Gases with Convertible Components

10.3 Conjoint Analysis and addals()

Chapter 11

Discriminant Analysis and criminals()

Placeholder

11.1 Equations

11.2 Examples

11.2.1 Iris data

Chapter 12

Multiblock Canonical Correlation and overals()

Placeholder

12.1 Equations

12.2 Examples

12.2.1 Thirteen Personality Scales

Chapter 13

Code

Placeholder

13.1 R Code

13.1.1 Driver

13.1.2 Engine

13.1.3 Aspect Engine

13.1.4 Some Aspects

13.1.5 Structures

13.1.6 Wrappers

13.1.7 Splines

13.1.8 Gram-Schmidt

13.1.9 Cone regression

13.1.10 Coding

13.1.11 Utilities

13.2 C Code

13.2.1 Splines

13.2.2 Gram-Schmidt

13.2.3 Coding

Chapter 14

Backmatter

14.1 Key Names and Symbols

14.2 References

De Leeuw, J., and P. Mair. 2009. “Homogeneity Analysis in R: the Package homals.” *Journal of Statistical Software* 31 (4): 1–21. https://www.jstatsoft.org/v31/i04/.

Gifi, A. 1980. *Niet-Lineaire Multivariate Analyse [Nonlinear Multivariate Analysis]*. Leiden, The Netherlands: Department of Data Theory FSW/RUL.

———. 1981. *Nonlinear Multivariate Analysis*. Leiden, The Netherlands: Department of Data Theory FSW/RUL.

———. 1990. *Nonlinear Multivariate Analysis*. New York, N.Y.: Wiley.

Hill, M. O. 1990. “Review of A. Gifi, Multivariate Analysis.” *Journal of Ecology* 78 (4): 1148–49.

IBM. 2015. *IBM SPSS Categories 23*. IBM Corporation.

Marvell, A. 1653. “The Character of Holland.”

Meulman, J. J., and W. J. Heiser. 2012. *IBM SPSS Categories 21*. IBM Corporation.

Michailidis, G., and J. De Leeuw. 1998. “The Gifi System for Descriptive Multivariate Analysis.” *Statistical Science* 13: 307–36.

SPSS. 1989. *SPSS Categories*. SPSS Inc.

Takane, Y. 1992. “Review of Albert Gifi, Nonlinear Multivariate Analysis.”

Journal of the American Statistical Association 87: 587–88.

Van der Heijden, P. G. M., and S. Van Buuren. 1916. “Looking Back at the Gifi System of Nonlinear Multivariate Analysis.” *Journal of Statistical Software* 73 (4).

Van der Velden, M. 2012. “On Generalized Canonical Correlation Analysis.” In *Proceedings 58th World Statistical Congress, 2011, Dublin*, 758–65. The Hague: International Statiatical Instutute.

Xie, Y. 2015. *Dynamic Documents with R and knitr.* Second Edition. CRC Press.

———. 2016. *Bookdown: Authoring Books with r Markdown.*