Journal of Statistical Software

September 2016, Volume 73, Issue 4. doi: 10.18637/jss.v073.i04

Looking Back at the Gifi System of Nonlinear Multivariate Analysis

Peter G. M. van der Heijden
Utrecht University
University of Southampton

Stef van Buuren
Utrecht University
TNO Leiden

Abstract

Gifi was the nom de plume for a group of researchers led by Jan de Leeuw at the University of Leiden. Between 1970 and 1990 the group produced a stream of theoretical papers and computer programs in the area of nonlinear multivariate analysis that were very innovative. In an informal way this paper discusses the so-called Gifi system of nonlinear multivariate analysis, that entails homogeneity analysis (which is closely related to multiple correspondence analysis) and generalizations. The history is discussed, giving attention to the scientific philosophy of this group, and links to machine learning are indicated.

Keywords: homogeneity analysis, multiple correspondence analysis, HOMALS, CATPCA.

1. Introduction

The late 70s and early 80s was an interesting time at the Faculty of Psychology at the University of Leiden. Professor John van de Geer, a cognitive psychologist who became interested in multivariate analysis and, among others, wrote an influential book on a geometrical approach to multivariate analysis (Van De Geer 1971), had succeeded to establish a group called "Data Theory" (compare Van Der Heijden and Sijtsma 1996; Heiser 2008). The group was located in a remote corner of the building, had an interfacultary status aparte that freed them up from regular teaching, and they were doing Big Science. According to Heiser (2008), Jan de Leeuw was the first person who Van de Geer recruited for his Data Theory department. The group combined a strong sense of geometric thinking with emerging computational tools into state-of-the-art quantitative data analysis techniques that has produced yet unseen visualizations of data by a series of computer programs with names like HOMALS, PRINCALS, CANALS, PRIMALS, ANACOR and others, written in Fortran, with experimental versions written in

APL. The group emphasised exploratory data analysis and optimization algorithms, and had a dislike for distributional assumptions. Of course, such free and refreshing radicals attracted a lot of attention from students like ourselves. The world would soon abandon old statistics and embrace the novel ways of data analysis. And we were part of that!

Besides a stream of theoretical papers and computer manuals, the main products of the group were Gifi (1980) (in Dutch) and Gifi (1981) (in English). These were not really books. Both were printed by the local repro in A4 format. Albert Gifi was not the name of a single author, but was a nom de plume for the following individuals: Bert Betonville, Steef de Bie, Eeke van der Burg, John van de Geer, Willem Heiser, Judy Knip, Jan de Leeuw, Jacqueline Meulman, Peter Neufeglise, André Nierop and Ineke Stoop. The preface of Gifi (1981) states: "The portrait ... 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 many years, to the cause he served." Gifi was the servant of Francis Galton for 40 years. When Galton died, a heritage of 45,000 pounds was given to the University of London and only 200 pounds to Albert Gifi (Heiser 2008).

The scientific father of the Gifi group was Jan de Leeuw. In fact, much of the material can be traced back to his dissertation (De Leeuw 1973, the thesis was written under the supervision of John van de Geer) and his early work with Forest Young and Yoshio Takane (Young 1981). After Jan moved to UCLA in 1987, Gifi (1981) was reworked by Willem Heiser, Jacqueline Meulman and Gerda van den Berg into what would become the book Gifi (1990) as published by John Wiley & Sons. A related Wiley book was published in 1988 (Van Rijckevorsel and De Leeuw 1988) and two related Sage books by John van de Geer in 1993 (Van De Geer 1993a,b), published in the series Advanced Quantitative Techniques in the Social Sciences Series, edited by Jan de Leeuw. A good overview of the Gifi system of nonlinear multivariate analysis can be found in Michailidis and De Leeuw (1998).

2. Technical description

The data analytic tool homogeneity analysis was the starting point of the Gifi system of nonlinear multivariate analysis. The Gifi system extends homogeneity analysis by placing various additional restrictions on the solution.

Homogeneity analysis starts with a set of categorical variables measured on n observations that may be measured on any measurement level, i.e., that can nominal, ordinal or interval-valued variables. Interest goes into an exploratory analysis of the relationships between these variables. If the variables would be measurements on continuous scales, principal component analysis (PCA) would be a good candidate to answer the question of interest. But now that the variables are categorical, possibly of an ordinal or nominal measurement level, PCA cannot be applied in a straightforward way. The solution offered in homogeneity analysis is that that each of the categories of these variables is scaled, such that the variables are quantified. Now that the variables are quantified, it is possible to calculate a correlation between each pair of variables. The criterion for this quantification of each of the variables is that the quantification is chosen such that the resulting first eigenvalue of the correlation matrix is maximized, in other words, the variables are quantified in such a way that they become as similar as possible. This yields a first component score for each of the n observations.

For later dimensions different quantifications are found for each of the variables. I.e., for the

second dimension a second quantification is found in such a way that the resulting second correlation matrix has the highest possible eigenvalue, under the restriction that the n-vector of first component scores is orthogonal to the n-vector of the second component scores. And so on for further dimensions.

Homogeneity analysis received its name from the property of the method that it yields homogeneous groups of objects and categories. The name of the first computer program, HOMALS, refers to homogeneity analysis by alternating least squares, where alternating least squares is the algorithm used to calculate the solution. At the end of the 1980s the IBM SPSS module **Categories** (compare Meulman, Heiser, and SPSS 1999) became available. `HOMALS` was one of the programs (later on, `HOMALS` was renamed into `MULTIPLE CORRESPONDENCE`). Much later **homals** was made available in R (De Leeuw and Mair 2009).

One of the key reasons of the name Gifi *system* is that many extensions were proposed. These extensions were (1) to restrict the quantifications of the categories of a variable to be nominal, ordinal, or have an interval level, (2) to quantify the variables only once, (3) to apply the methodology on sets of variables, thus generalizing k-sets analysis for quantitative variables, (4) to apply the methodology on two sets of variables, yielding a generalization of canonical correlation analysis for categorial variables, and later (5) to use splines to transform continuous variables, (6) to generalize multiple regression, (7) to find groups of similar objects, and (8) to find optimal data transformations for dynamical systems and time series. In the Gifi system as originally developed in Leiden this led to programs with names like (1) PRINCALS, (2) PRIMALS, (3) OVERALS, (4) CANALS, (5) SPLINALS, (6) MORALS and (7) GROUPALS. In the IBM SPSS module **CATPCA** some of more important options sketched here are brought together. The basis of **CATPCA** has been PRINCALS. It was extended to deal with monotone splines, weighting options, different data preparation options, and so on.

With regard to the measurement levels chosen, it is interesting to note that, contrary to what one often encounters in the literature, a variable does not "have" an ordinal measument level because the categories are ordered, but it is the researcher who decides that in a specific analysis he wants to investigate a relationship between variables where for this variable ordinal transformations are allowed. In a follow-up analysis he can decide to treat the ordinal categories using nominal transformations or an interval transformation. This is a feature of the Gifi system that confuses novice users.

The original Gifi system has a few other remarkable properties that make it distinct from mainstream statistics. A first property is that it is a system for categorical data, not for continuous data. A second property is that it is very different from the standard approach to the analysis of categorical data proposed around that time by Bishop, Fienberg, and Holland (1975), which uses a likelihood-based statistical model starting from multinomial and Poisson distributions. Third, the tools in the Gifi system were defined from the start in terms of loss functions that had to be minimized (compare Husson, Josse, and Saporta 2016). In this sense the presentation was different from the standard approach of presenting statistical models, where first there is a model formulation (for example, a regression equation) and then a presentation of a least squares criterion or a likelihood that has to be minimized or maximized. Fourth, emphasis was given to the numerical algorithm to maximize the loss function, namely the alternating least squares algorithm. Fifth, the Gifi system was presented with a scientific philosophy making strong statements about the preferability of the exploratory data analytic approach taken, and the absurdity of the approach of taken by formulating a statistical model with its assumptions that, Gifi claimed, were unrealistic.

3. Scientific philosophy

The Gifi framework is inspired by Tukey's manifesto, "The future of data analysis" (Tukey 1962) and Benzécri (1973a,b) principle of II: "The model must follow the data, not the other way around." Gifi emphasizes techniques over models, and rejects the classical notion of statistical inference in which hypotheses are tested. Gifi (1990), p. 19, writes: "The data analytic approach does not start with a model, but looks for transformations and combinations of the variables with the explicit purpose of representing the data in a simple and comprehensive, and usually graphical, way."

While Gifi's preference for exploratory data analysis, computational techniques and visualization may now sound familiar, this certainly wasn't the case in the 70s and 80s. In fact, when it was published the Gifi book did not generate a lot of attention, and was well outside of mainstream statistics. The extreme position of the Leiden school has not gone unnoticed though. In 1987, the Dutch Association of Statistics (VVS-OR) organised a scientific debate under the name of "The formal approach from Groningen versus the informal approach from Leiden." The proponent of the Groningen school was Ivo Molenaar whereas the Leiden school of thought was represented by Jan de Leeuw. This debate has been documented in Molenaar (1988), De Leeuw (1988), and De Gruijter (1988).

Molenaar argued that the Gifi approach risks encouragement of a lazy attitude of "pick whatever data, use a Gifi technique, give some post hoc thoughts, and your research report is ready." Instead the researcher should consider the question "What would happen if one did it again?", which translates into dealing with issues like the selection of observational units, selection of variables and measurement procedures, measurement errors, assignment to treatments, chance capitalization and variation with time before the data are collected. Molenaar rejects Benzécri's principle 2: "The data have no voice, and only will speak and be understood when a skilful investigator has set up the draft of a design, an analysis and a model, based on the available knowledge."

In response, De Leeuw pointed out that there is an almost universal violation of the basic rules: "It is not true that people first formulate a model, then collect data, and then perform statistics." Decisions made by the scientist cannot be formalized before the data are collected. Moreover, the assumptions typically made in the standard statistical models do not make much sense. In many social and behavioral science applications the idea of independent replications is irrelevant. If it is impossible to replicate the experiment, then the idea of repeated experiments does not make sense either. De Leeuw describes inferential statistics as "a confusing and quite nonsensical collection of rituals." On the other hand, De Leeuw says, many of the techniques work, quite beautifully, for data analysis.

4. Reception and the future

It is a remarkable coincidence that in the 70s when homogeneity analysis became popular it was invented under different names in at least three places of the world: Nishisato in Canada presented dual scaling (see a summary in Nishisato 1981), Benzécri in France presented (multiple) correspondence analysis (see a summary in the books Benzécri 1973a,b) and the group of Jan de Leeuw presented homogeneity analysis. Nishisato, Benzécri and De Leeuw in turn made use of earlier work by the Japanese researcher Hayashi. Where each of these approaches had its own emphases on specific aspects of the methodology (the above

sketch of homogeneity is certainly not the only way to present the common methodology) these persons influenced each other. The work of Benzécri was originally written in French (both with regard to the language and in terms of mathematical generality), and was therefore only accessible by a few. However, the ideas of Benzécri were spread later in English by Michael Greenacre (Greenacre 1984), who popularized simple correspondence analysis (i.e., homogeneity analysis of two variables) and multiple correspondence analysis. Nishisato and the members of the Gifi group regularly met at meetings of the Psychometric Society.

What was the impact of the original Gifi system? This question is not easy to answer. An author impact analysis for Gifi, using the program Publish or Perish (Harzing 2007), finds about 2,250 citations for Gifi. But then citations from the 80s are likely to be missed, and impact should also be studied for authors linked to the Gifi project, not in the least, of course, Jan de Leeuw himself. Quite a task that we are happy to leave to someone else. However, we notice that, where nowadays researchers put most of their energy into writing manuscripts that can be published in journals, in the Netherlands around 1980, the climate was just changing from a situation where you could write internal reports that you sent to your close collegues in other universities, to more emphasis on publishing in journals. Publishing papers in journals was not as usual yet as it is now, and the members of the Gifi group wrote much of their work in internal reports that they sent to collegues they knew they were interested. We believe that if in this early period more papers would have appeared in journals, and Gifi (1981) would have been published straight away by an official publisher as a book, the impact would have been larger.

In comparison, Greenacre (1984) received over 4,000 citations. The book reads very well, and the focus is on a single element from the Gifi system, namely homogeneity analysis (that he called, following Benzécri, multiple correspondence analysis). Greenacre published in American and Anglo-Saxon statistical journals, produced several more applied books about correspondence analysis that were much cited, and organized conferences solely devoted to multiple correspondence analysis. This helped to increase the popularity of homogeneity analysis/multiple correspondence analysis enormously. Around the same time, ideas and techniques closely related to Gifi were put into practice in a regression context, leading up to successful methodologies like the ACE algorithm (Breiman and Friedman 1985) and generalized additive models (Hastie and Tibshirani 1990).

The Gifi group was able to get important parts of the Gifi system into the computer package IBM SPSS, and also other computer packages have routines for (multiple) correspondence analysis. This made it much easier for users to apply the Gifi system. So in all, whereas it is a reasonable question to ask what the impact of the Gifi group was, it is hard to give a definite answer.

Interestingly, the way in which the methodology is presented in the Gifi system, namely through loss functions, is similar to the way methodology is sometimes presented in machine learning. In machine learning criteria are regularly specified that are optimized directly, without making use of traditional statistical models being formulated in, for example, a likelihood framework. Thus the Gifi system fits well in the machine learning arena, was ahead of its time, and it has the potential to have "a second life" in the analysis of big data.

The question is, though, who will be able to create this second life? Of course, computational techniques have changed tremendously over the last 30 years, so one would do calculations in a different way now. We believe, however, that the Gifi system is still modern in its way

of specifying sensible constraints on the solution, in its approach to create low-dimensional geometric visualizations of large data sets, and in its focus on exploratory multivariate analysis. We would encourage new generations of data scientists to browse an old copy of the 1981 or 1980 Gifi books, or visit a modernized version (De Leeuw, Mair, and Groenen 2016). Hopefully, they will enjoy Gifi's spirit as much as we did.

5. Personal

Jan de Leeuw is a charismatic leader and very sympathetic towards his students. We were both Ph.D. students of Jan. After Jan moved to Los Angeles in 1986, there was a substantial stream of Dutch visitors to UCLA. In UCLA, Jan had two rooms available in a corner of the huge department headed by Peter Bentler in Franz Hall. When we visited him, he would take the office space without windows and offer us the office space with windows. Edith de Leeuw and Joop Hox were even kindly offered Jan's private home up in the hills as the perfect place to celebrate their marriage. Other visitors around that time included Catrien Bijleveld, Patrick Groenen, Paul Bekker, Jan van Rijckevorsel and Kees van Montfort, who – fully unprepared – ran and finished the 1988 LA marathon. Together with his former Leiden Ph.D. students like Willem Heiser, Jacqueline Meulman and Peter Kroonenberg all have become professors in The Netherlands, thus demonstrating Jan's lasting contribution on the Dutch academic world of methodology and statistics for the social sciences.

References

Benzécri JP (1973a). *L'Analyse Des Données: Tome I: La Taxinomie*. Dunod, Paris.

Benzécri JP (1973b). *L'Analyse Des Données: Tome II: L'Analyse Des Correspondances*. Dunod, Paris.

Bishop YMM, Fienberg SE, Holland PW (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge.

Breiman L, Friedman JH (1985). "Estimating Optimal Transformations for Multiple Regression and Correlation." *Journal of the American Statistical Association*, **80**(391), 580–598. doi:10.1080/01621459.1985.10478157.

De Gruijter DNM (1988). "Data Analysis and Statistics, Report of a Discussion." *Statistica Neerlandica*, **42**(2), 99–102. doi:10.1111/j.1467-9574.1988.tb01223.x.

De Leeuw J (1973). *Canonical Analysis of Categorical Data*. Ph.D. thesis, Leiden University: Psychological Institute. Republished in 1984 by DSWO Press, Leiden.

De Leeuw J (1988). "Models and Techniques." *Statistica Neerlandica*, **42**(2), 91–98. doi:10.1111/j.1467-9574.1988.tb01222.x.

De Leeuw J, Mair P (2009). "Gifi Methods for Optimal Scaling in R: The Package **homals**." *Journal of Statistical Software*, **31**(1), 1–21. doi:10.18637/jss.v031.i04.

De Leeuw J, Mair P, Groenen P (2016). "Gifi Analysis of Multivariate Data." Unpublished book draft, URL http://gifi.stat.ucla.edu/gifi/_book/.

Gifi A (1980). "Niet-Lineaire Multivariate Analyse." *Technical report*, University of Leiden, Department of Data Theory, Leiden.

Gifi A (1981). "Nonlinear Multivariate Analysis." *Technical report*, University of Leiden, Department of Data Theory, Leiden.

Gifi A (1990). *Nonlinear Multivariate Analysis*. John Wiley & Sons, New York.

Greenacre M (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Harzing AW (2007). "Publish or Perish." URL http://www.harzing.com/pop.htm.

Hastie T, Tibshirani R (1990). *Generalized Additive Models*. Chapman and Hall, London.

Heiser WJ (2008). "Psychometric Roots of Multidimensional Data Analysis in the Netherlands: From Gerard Heymans to John van de Geer." *Electronic Journ@l for History of Probability and Statistics*, **4**(2). URL http://www.jehps.net/Decembre2008/Heiser.pdf.

Husson F, Josse J, Saporta G (2016). "Jan De Leeuw and the French School of Data Analysis." *Journal of Statistical Software*, **73**(6), 1–17. doi:10.18637/jss.v073.i06.

Meulman JJ, Heiser WJ, SPSS (1999). *SPSS Categories 10.0*. SPSS, Chicago.

Michailidis G, De Leeuw J (1998). "The Gifi System of Descriptive Multivariate Analysis." *Statistical Science*, **13**(4), 307–336. doi:10.1214/ss/1028905828.

Molenaar IW (1988). "Formal Statistics and Informal Data Analysis, or Why Laziness Should Be Discouraged." *Statistica Neerlandica*, **42**(2), 83–90. doi:10.1111/j.1467-9574.1988.tb01221.x.

Nishisato S (1981). *Analysis of Categorical Data: Dual Scaling and Its Applications*. University of Toronto Press, Toronto.

Tukey JW (1962). "The Future of Data Analysis." *The Annals of Mathematical Statistics*, **33**(1), 1–67. doi:10.1214/aoms/1177704711.

Van De Geer JP (1971). *Introduction to Multivariate Analysis for the Social Sciences*. Freeman and Company, San Francisco.

Van De Geer JP (1993a). *Multivariate Analysis of Categorical Data: Applications*. Advanced Quantitative Techniques in the Social Sciences. Sage Publications, Newbury Park.

Van De Geer JP (1993b). *Multivariate Analysis of Categorical Data: Theory*. Advanced Quantitative Techniques in the Social Sciences. Sage Publications, Newbury Park.

Van Der Heijden PGM, Sijtsma K (1996). "Fifty Years of Measurement and Scaling in the Dutch Social Sciences." *Statistica Neerlandica*, **50**(1), 111–135. doi:10.1111/j.1467-9574.1996.tb01483.x.

Van Rijckevorsel JLA, De Leeuw J (eds.) (1988). *Component and Correspondence Analysis*. John Wiley & Sons.

Young FW (1981). "Quantitative Analysis of Qualitative Data." *Psychometrika*, **46**(4), 357–388. doi:10.1007/bf02293796.

Affiliation:

Peter G.M. van der Heijden
Department of Methodology and Statistics
Utrecht University
P.O. Box 80.140
3508 TC Utrecht, The Netherlands
E-mail: P.G.M.vanderHeijden@uu.nl
URL: http://www.uu.nl/staff/PGMvanderHeijden/

Stef van Buuren
Department of Methodology & Statistics
University of Utrecht
P.O. Box 80140
3508 TC Utrecht, The Netherlands
E-mail: Stef.vanBuuren@tno.nl
URL: http://www.stefvanbuuren.nl/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
September 2016, Volume 73, Issue 4 *Submitted:* 2016-03-24
doi:10.18637/jss.v073.i04 *Accepted:* 2016-06-01