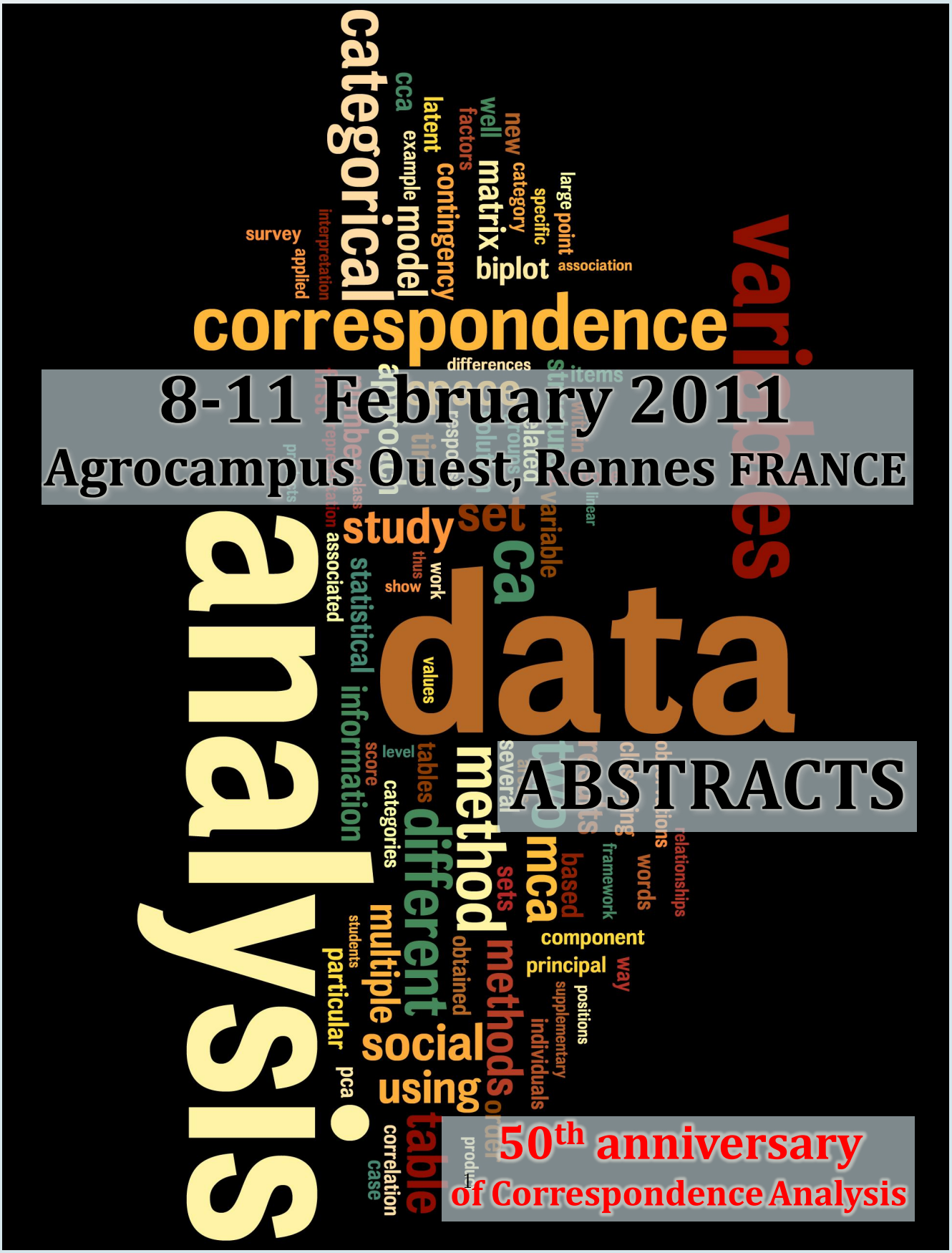
AGRO CAMPUS OUEST

The 6th CARME conference

Correspondence Analysis and Related MEthods



Yuna Blum design

International Conference on

CORRESPONDENCE ANALYSIS AND RELATED METHODS

(CARME 2011)

Agrocampus Ouest
Rennes

8 – 11 February 2011

Scientific Committee

- Jörg Blasius (University of Bonn)
- Michael Greenacre (Universitat Pompeu Fabra)
- Jérôme Pagès (Agrocampus Rennes)

Local Organizing Committee

- Julie Josse
- Karine Bagory
- Yuna Blum
- Marine Cadoret
- François Husson
- Sébastien Lê
- Vinciane Marchais
- Alain Bernardeau
- Anne Bourdeau
- Stéphane Crespel
- Gabriel Jalam
- Elisabeth Lenauld
- Jérôme Pagès

Jean-Paul Benzécri, 1st April 2009,
Symposium on Statistical Learning and Data Science, University Paris Dauphine
(snapshot Jérôme Pagès).

Cover design
Graphism: Michael Greenacre. The word cloud on the cover design is based on the words of all the abstracts of this CARME conference, so in a certain sense it is an "analysis" (most frequent words, and size related to frequency)
Design: Yuna Blum

Avant-propos

En 1960 Jean-Paul Benzécri fut nommé professeur à Rennes. En 1963, il donna son premier exposé sur l'Analyse des Correspondances. Dans l'intervalle, cette singulière méthode était née : ce congrès fête le cinquantième anniversaire de cette naissance.

Rapidement, une équipe l'entoura. Au premier rang de laquelle, Brigitte Escofier, élève de la première heure et contributrice décisive dès cette époque puisque c'est à ce moment qu'elle découvrit les relations de transition. Que l'on y songe un instant : en 1961, l'analyse en composantes principales existait depuis presque trente ans et ces relations, qui, *mutatis mutandis*, s'y appliquent aussi, étaient passées inaperçues ! Par la suite, pendant trente années, Brigitte a enrichi l'analyse des données par de nombreuses recherches importantes (rassemblées en 2003 dans un livre) et tout naturellement un congrès à Rennes sur l'analyse des correspondances ne peut qu'être aussi en son hommage.

Outre Brigitte Escofier, citons deux autres élèves de la première heure : Marie-Odile Lebeaux et Brigitte Le Roux -Brigitte nous parlera de la période rennaise. En septembre 1965, Jean-Paul Benzécri partit pour la capitale avec son équipe.

A la fin des années 60, Michel Kerbaol, ancien de l'équipe parisienne, revint sur Rennes, où, travaillant à l'INSERM, il introduisit l'analyse des correspondances dans le monde médical. C'est précisément Michel Kerbaol qui fit ce qui fut sans doute le premier cours d'analyse des données sur Rennes. C'était en 1971, à Agrocampus. Il suscita l'enthousiasme de plusieurs étudiants. L'un d'entre eux, votre serviteur, devenu enseignant l'année d'après, y développa, autour de l'analyse des correspondances, un enseignement d'abord, une recherche ensuite, solidement épaulé en cela par l'équipe fondatrice, alors parisienne : c'est ainsi que, par ordre alphabétique, Pierre Cazes, Ludovic Lebart et Maurice Roux vinrent à plusieurs reprises nous apporter la bonne parole. Nous sommes enchantés de les accueillir à nouveau à l'occasion de ce congrès.

Aujourd'hui, l'analyse des données est une composante principale d'Agrocampus. En enseignement, tous les étudiants reçoivent un cours de base et les plus chanceux bénéficient, en seconde année de master, d'un cours approfondi. Ces derniers participent activement à ce congrès, sur le plan matériel en aidant à l'organisation et sur le plan scientifique en présentant, sous forme de posters, un travail personnel réalisé ces deux derniers mois au laboratoire de mathématiques appliquées. En recherche, plusieurs nouvelles méthodes, en particulier dans le domaine des tableaux multiples, ont été mises au point à Agrocampus. Elles sont rassemblées dans le logiciel libre FactoMiner, package R qui contient aussi les méthodes classiques. L'importance des package R n'est plus à démontrer : la première journée de ce congrès leur est dédiée.

A l'origine, l'analyse des correspondances fut mise au point pour ... analyser des données. Et, fait singulier en statistique, sa diffusion doit beaucoup aux utilisateurs. De ce point de vue, ce congrès CARME est bien dans l'esprit originel : bon nombre de sessions ont comme dénominateur commun, et donc comme intitulé, un domaine d'application : parmi elles, une place prépondérante revient aux sciences sociales.

Ce n'est pas un hasard : les sciences sociales sont à la fois un domaine qui nous concerne tous et

un terrain de prédilection pour l'analyse des correspondances. Ajoutons que l'initiative, en 1991, en revient à Walter Kristof, alors à l'institut de sociologie de Hambourg qui, avec Jörg Blasius and Michael Greenacre organisèrent le premier CARME en 1991 à Cologne. Du fait du succès de cette conférence, Michael and Jörg ont assuré la continuité de ces conférences qui se sont déroulées à Cologne (Correspondence Analysis in the Social Sciences 1991, Visualization of Categorical Data 1995, Large Scale Data Analysis 1999), à Barcelone (CARME 2003) et à Rotterdam (CARME 2007). Nous commémorons donc aussi les 20 années de CARME.
Au moment où ces lignes sont écrites, en janvier 2011, d'un certain point de vue rien ne s'est encore passé. Et pourtant, que d'énergie a été dépensée pour que ce congrès puisse avoir lieu ! Je ne saurais trop remercier :

- la direction d'Agrocampus, en la personne de Grégoire Thomas, Directeur général, qui a toujours soutenu la statistique en tant que discipline à part entière (et non uniquement au service des autres, ce qui ne va pas nécessairement de soi dans une institution centrée sur un domaine d'application) ;

- le pack logistique d'Agrocampus : Anne Bourdeau (questions administratives), Stéphane Crespel (questions pratiques), Vinciane Marchais (inscriptions en ligne), Alain Bernardeau (questions administratives non euclidiennes) ; sans eux, rien n'est possible.

- le laboratoire de mathématiques appliquées : Karine Bagory, Marine Cadoret, David Causeur, Thibaut Dutrion, Magalie Houée, François Husson, Julie Josse, Maela Kloareg, Sébastien Lê, Elisabeth Lenauld, Marie Verbanck ; là est le cœur du cœur ;

- plusieurs amis au sein d'Agrocampus : Gabriel Jalam (films), Yuna Blum (design) ; pour les " plus " indispensables.

Jérôme Pagès
14 janvier 2011

Brigitte Escofier (cliché Jean-Pierre Escofier).

Foreword

In 1960, Jean-Paul Benzécri was appointed Professor in Rennes. In 1963, he gave the first lecture on Correspondence Analysis. In the meantime, this singular approach to data analysis was born: this conference celebrates its fiftieth birthday.

Quickly, a team surrounded him. At the forefront was Brigitte Escofier, one of the first students and to become a decisive contributor. It is at this time that she discovered the transition formulas. Let's think a moment about this: in 1961, principal component analysis has existed for almost thirty years and these formulas, mutatis mutandis applicable to PCA, were unnoticed! Subsequently, during thirty years, Brigitte enriched data analysis by many important researches (gathered in a book in 2003); naturally, a conference in Rennes on correspondence analysis is necessarily a tribute to her too.

In addition to Brigitte Escofier, let's remember two other students of that first period: Marie-Odile Lebeaux and Brigitte Le Roux - Brigitte will talk about the "Rennes period" at our conference. In September 1965, Jean-Paul Benzécri went to Paris with his team.

In the late '60s, Michel Kerbaol, formerly member of the Paris team, returned to Rennes; working in INSERM, he introduced correspondence analysis in the medical world. It is precisely Michel Kerbaol who gave what was probably the first data analysis lecture in Rennes, at Agrocampus in 1971. He aroused the enthusiasm of several students. One of them, yours truly, became assistant-professor one year later: he developed, around correspondence analysis, first a course then research, securely helped by the founders, who were working then in Paris. So, alphabetically, Pierre Cazes, Ludovic Lebart and Maurice Roux came several times to Rennes to bring us the good word. We are delighted to welcome them again in this Conference.

Today, data analysis is a principal component of Agrocampus. All the students attend a basic course and the luckiest ones, in the second year of their master degree, a comprehensive course. These last students actively participate in this Conference, practically by helping organization, and scientifically, by the way of posters, fruit of their personal work done these last two months in the Applied Mathematics department. Several new methods, in particular in the field of multiple tables, were developed at Agrocampus. They are gathered in the free software FactoMiner, an R package that also contains the classical methods. The importance of the R packages is now evident: the first day of this conference is devoted to them.

Originally, correspondence analysis was developed for ... analysing data! And its dissemination owes much to users. From this point of view, this CARME Conference is in the original spirit: many sessions have, as a common denominator, and therefore as a title, an application domain, and among them a prominent place is devoted to the social sciences.

It is not a coincidence: the social sciences are both an area which concerns all of us and a marvellous field for correspondence analysis. In fact, in 1991, CARME was initiated by Walter Kristof, at this time Professor at the Institute of Sociology of Hamburg, with Jörg Blasius and Michael Greenacre organizing the first meeting in 1991 in Cologne. Because of the success of this conference, Michael and Jörg have assured the continuity of the CARME Conferences, which have taken place three

times in Cologne (Correspondence Analysis in the Social Sciences 1991, Visualization of Categorical Data 1995, Large Scale Data Analysis 1999), then in Barcelona (CARME 2003) and Rotterdam (CARME 2007). So we also celebrate 20 years of CARME at our conference in Rennes.
At the time these lines are written, in January 2011, from a certain point of view nothing has happened. But what an amount of energy has been spent so that this Conference can take place! It's nice to thank:

- the direction of Agrocampus, in the person of Grégoire Thomas, Director General, who always supports statistics as a discipline in its own right (and not only in the service of others, which is not necessarily evident in an institution centred on an application field).
- Agrocampus logistic pack: Anne Bourdeau (administrative matters), Stéphane Crespel (practical issues), Vinciane Marchais (online registration), Alain Bernardeau (non-Euclidean administrative matters); without them, nothing is possible.
- Applied Mathematics department: Karine Bagory, Marine Cadoret, David Causeur, Thibaut Dutrion, Magalie Houée, François Husson, Julie Josse, Sébastien Lê, Elisabeth Lenauld, Marie Verbanck ; the heart of the heart.
- several friends in Agrocampus: Gabriel Jalam (movies), Yuna Blum (design); the indispensable "plus" factor.

Jérôme Pagès
14 January 2011

Michael Greenacre, Walter Kristof (who initiated the original CARME conference in 1991) and Jörg Blasius.

Contents

Wednesday 9 February **13**

09:30 - 10:30 Invited speaker (Room Matagrin, chair M. Greenacre) **13**

Volpato Richard: *Letting data speak: Enunciative modalities of correspondence analysis* 13

Le Roux Brigitte: *Birth of CA in Rennes and what's new 40 years later?* 14

11:00 - 13:00 CA & MCA: Theory & algorithms (Room Rieffel, chair M. Friendly) **15**

Greenacre Michael: *Unifying the geometry of simple and multiple correspondence analysis* 15

Choulakian Vartan, De Tibeiro Jules: *Graph partitioning by correspondence analysis and taxicab correspondence analysis* 16

D'Ambra Luigi, Beh Eric, Camminatiello Ida: *Singly and doubly ordered cumulative correspondence analysis* 17

Josse Julie, Chavent Marie, Liquet Benoît, Husson François: *Handling missing values with regularized iterative multiple correspondence analysis* 18

Langovaya Anna, Kuhnt Sonja: *Correspondence analysis and moderate outliers* 19

Séguéla Julie, Saporta Gilbert: *A comparison between latent semantic analysis and correspondence analysis* 20

11:00 - 13:00 Related methods: Applications (Room 1, chair A. Mom) **21**

Bougeard Stéphanie, Qannari El Mostafa, Fablet Christelle: *Multiblock method for categorical variables. Application to air quality in pig farms* 21

Verbanck Marie, Lê Sébastien, Pagès Jérôme: *Towards the integration of biological knowledge with canonical correspondence analysis when analyzing Xomic data in an exploratory framework* 22

Le Pouliquen Marc, Csernel Marc: *Betweenness relation orientated by Guttman effect in critical edition* 23

Weisz Robert, Karim Jahanvash: *Weisz communication styles inventory (WCSI-version 1.0): Development and validation* 24

Stanimir Agnieszka, Grzeskowiak Alicja, Dziechciarz Jozef: *Application of correspondence analysis and related methods in evaluation of knowledge and skills of young Peo* . . 25

Korneliussen Tor: *Information sources that EU tourists use: A cross-country study* . . . 26

14:30 - 16:00 Invited speaker (Room Matagrin, chair J. Pagès) **27**

Lebart Ludovic: *About the history of multiple correspondence analysis* 27

Cazes Pierre: *Some comments on correspondence analysis* 28

Roux Maurice: *Cluster analysis with k-means: what about the details ?* 29

16:00 - 16:20 Speed presentation of posters (Room Rieffel, chair J. Pagès) **30**
Böcük Harun, Asan Zerrin, Türe Cengiz: *The chemical analysis of soil ? Plant with high boron concentrations by log-ratio analysis* . . . 30
Durucasu Hasan, Ican Özgür: *Evaluation of Turkish media and the athletics news by correspondence analysis* . . . 31
Fersi Kmar, Benlagha Noureddine, Ben Ammou Samir: *The insurability of risks: A quantitative approach applied to motor insurance* . . . 32
Nenadic Oleg, Greenacre Michael: *Correspondence Analysis in R: the ca package* . . . 33
Palacios Fenech Javier: *Principal component analysis of international diffusion of durable goods* . . . 34
Souza Marcio, Bastos Ronaldo, Vieira Marcel: *The derivation of individual overall attitude scores from a multiple correspondence analysis solution* . . . 35
de Tibeiro Jules, Murdoch Duncan: *Correspondence analysis with incomplete paired data using Bayesian analysis* . . . 36
Wurzer Marcus, Mair Patrick: *Gifi methods to explore EU-SILC data* . . . 37

17:00 - 18:20 Related methods: Computation (Room Rieffel, chair P. Mair) **38**
Dossou-Gbété Simplice: *Analyzing multiple time series using a dynamic latent variables principal component analysis model* . . . 38
Iodice D'Enza Alfonso, Palumbo Francesco: *An evolutionary analysis of association patterns* . . . 39
Vehkalahti Kimmo, Sund Reijo: *First 50 years of Survo: from a statistical program to an interactive environment for data processing* . . . 40
Morin Annie, Pham Nguyen-Khang: *Interactive image mining* . . . 41

17:00 - 18:20 Sensory analysis (Room 1, chair M. Bécue-Bertaut) **42**
Buche Marianne, Cadoret Marine, Lê Sébastien: *Projective tests using Napping, the Rorschach test revisited: are the cultural differences between Asians and Caucasians significant?* . . . 42
Cadoret Marine, Lê Sébastien, Pagès Jérôme: *Euclidean representations of a set of hierarchies using multiple factor analysis* . . . 43
Qannari Mostafa, Courcoux Philippe, Cariou Véronique: *Analysis of sorting data using multiple correspondence analysis and a related method* . . . 44
Worch Thierry, Lê Sébastien, Pagès Jérôme: *Validation of ideal profile data using multivariate analysis: The ideal products? Space as a link between the products and their preferences* . . . 45

Thursday 10 February **46**

09:00 - 10:30 Invited speaker (Room Matagrin, chair L. Lebart) **46**
Bécue-Bertaut Mónica: *Textual and lexical statistics* . . . 46
Kroonenberg Pieter: *Three-mode correspondence analysis: Some history and an ecological example from the sea bed* . . . 47

11:20 - 12:00 Three-Way Data (Room Rieffel, chair P. Kroonenberg) **48**
Bénasséni Jacques, Bennani Dosse Mohammed: *The power STATIS-ACT method* . . . 48
Cadot Martine, Lelu Alain: *Representing interaction in multiway contingency tables: MIDOVA, CA and log-linear model* . . . 49

12:00 - 13:00 Textual Data Analysis (Room Rieffel, chair L. d'Ambra) 50
Kostov Belchin Adriyanov, Bécue-Bertaut Mónica, Morin Annie: *Canonical correspondence analysis for uncovering temporal features in chronological textual data* 50
Hornbostel Stefan, Marty Christoph: *Excellent news for German universities? A multiple correspondence analysis of media reporting on the excellence initiative* 51
Cadoret Marine, Buche Marianne, Lê Sébastien: *Confidence ellipses when analyzing simultaneously several contingency tables resulting from free-text descriptions* 52

11:00 - 13:00 MCA-methods in the social sciences (Room 1, chair F. Murtagh) 53
Le Roux Brigitte, Bienaise Solène: *Combinatorial inference in geometric data analysis: typicality test* 53
Blasius Jörg: *Screening the data for detecting methodological induced variation* 54
Souza Augusto, Bastos Ronaldo, Vieira Marcel: *Complex sampling designs and multiple correspondence analysis* 55
Zárraga Amaya, Goitisolo Beatriz: *Correspondence analysis of surveys with conditioned and multiple response questions* 56
Lubbe Sugnet, Silal Sheetal, Niel le Roux: *Constructing a socio-economic status index for a non-homogeneous society with distinct sets of variables in multiple correspondence analysis* 57
Grannell Andrew, Fitzgerald Tony, Corcoran Paul: *Deliberate self harm among Irish adolescents* 58

14:30 - 16:00 Invited speaker (Room Matagrin, chair G. Saporta) 59
Gower John: *Biplots: Taking stock* 59
Friendly Michael, Turner Heather, Firth David, Zeileis Achim: *Advances in visualizing categorical data* 60

16:30 - 17:50 Biplots (Room Rieffel, chair J. Gower) 61
De Rooij Mark: *The mixed effect trend vector model* 61
Graffelman Jan: *New pictures for correlation structure* 62
Vicente-Villardon Jose Luis: *Logistic biplots for binary, nominal and ordinal data* . . . 63
Vines Karen: *Predictive nonlinear biplots: maps and trajectories* 64

17:50 - 18:30 MFA: Applications (Room Rieffel, chair S. Bougeard) 65
Morand Elisabeth, Garnier Bénédicte, Bonvalet Catherine: *Multiple factor analysis to two-way contingency table to compare residential and geographical trajectories* . . . 65
Ganón Elena: *Simultaneous analysis of contingency tables drawn with telephone data registration from the national telephone service to support women suffering violence in Uruguay* 66

16:30 - 18:30 Social space: Methodology (Room 1, chair B. Le Roux) 67
Murtagh Fionn, Ganz Adam, Reddington Joe: *Semantics of narrative in collective, distributed problem-solving environments based on correspondence analysis and hierarchical clustering* 67
Hjellbrekke Johs., Korsnes Olav: *Cultural distinctions: A geometric data analysis* . . . 68
Bernard Françoise, Goldfarb Bernard, Pardoux Catherine, Touati Myriam, Summa-Gettler Mireille: *Evaluation of seminars by correspondence analysis and related methods* 69

Ekelund Bo, Börjesson Mikael: *Mapping a citational universe: A GDA of literary dissertation bibliographies* . . . 70
Mühlichen Andreas: *Nominal, ordinal and metric variables in the "social space" ? Using CatPCA to examine lifestyles and regional identities in a medium-sized German city* 71
Frederiksen Jan Thorhauge: *Cross-over methodologies: correspondence analysis as a framework for mixed methods.* . . . 72

Friday 11 February **73**

09:00 - 10:30 Invited speaker (Room Matagrin, chair D. Causeur) **73**
ter Braak Cajo: *History of canonical correspondence analysis (CCA) in ecology* . . . 73
Dray Stéphane: *Analyzing spatial multivariate structures* . . . 74

11:00 - 12:00 Modelling (Room Rieffel, chair J. Bénasséni) **75**
Dossou-Gbété Simplice, Falguerolles Antoine de: *The Poisson trick for matched tables: a case for putting the fish in the bowl* . . . 75
Lombardo Rosaria, Beh Eric: *The aggregate prediction index and non-symmetrical correspondence analysis of aggregate data: The 2x2 table* . . . 76
Tenenhaus Michel, Tenenhaus Arthur: *Regularized generalized canonical correlation analysis* . . . 77

12:00 - 13:00 Clustering (Room Rieffel, chair M. Roux) **78**
Fernández-Aguirre Karmele, Garín-Martín Maria Araceli, Modroño-Herrán Juan Ignacio: *Visual displays. Some evidence through artificial and real data* . . . 78
Markos Angelos, Menexes George: *Hierarchical clustering on special manifolds* . . . 79
Tortora Cristina, Palumbo Francesco, Gettler Summa Mireille: *CD-clustering* . . . 80

11:00 - 13:00 Social space: Application (Room 1, chair J. Blasius) **81**
Funnell Robert: *Urban aboriginal lifestyles in Brisbane: mapping vertical and lateral stratification of opportunity for marginalised groups* . . . 81
Rosenlund Lennart: *Social and spatial structures in an urban environment* . . . 82
Bonnet Philippe, Lebaron Frédéric: *Latest methodological breakthroughs in geometric data analysis of cultural practices* . . . 83
Börjesson Mikael, Melldahl Andreas: *The Swedish social space of 1990. Investigating its structure and history* . . . 84
Lidegran Ida, Palme Mikael: *Out-of-study practices and symbolic capital among Swedish students in higher education* . . . 85
Frederiksen Morten: *Not so trustful after all? A study of trust, tolerance and solidarity in Denmark* . . . 86

14:30 - 16:00 Invited speaker (Room Matagrin, chair A. de Falguerolles) **87**
de Leeuw Jan: *History of Nonlinear Principal Component Analysis* . . . 87
Groenen Patrick: *Past, present, and future of multidimensional scaling* . . . 88

Author index **89**

Letting data speak: Enunciative Modalities of Correspondence Analysis

Richard Volpato
Manager,
Data Quality and Analysis,
Copyright Agency Ltd, Australia.
Richard@Volpato.net

Keywords: Correspondence Analysis, Visualization, Verbalization, Duality

After a fateful meeting with J.P. Benzecri in Paris in the mid 80s, while returning to Australia from Cambridge, the full force of Correspondence Analysis dawned upon me: well might it reduce multivariate data and tables to efficient summaries and compelling displays, but its real power to me was how it provoke original thoughts mediated by data. Practically this becomes a quest to convert visualizations to "verbalizations"; to produce a rhetoric (in the classical sense) of data. So from the simplicity of a proportion producing a sentence; to a three-way table producing several paragraphs; to dimensions prompting "names" of underlying factors and the whole "duality-diagram" providing a way by which expertise on a subject (inter-variable relations) can be mapped into on-the-ground experiences of the same subject (inter-object relations). Whole modalities of speaking come into alignment hitherto not disciplined by a rationality deep enough to encompass both. A matrix of data can, once rendered as a space, can reveal destinies prompting discourses never before imagined.

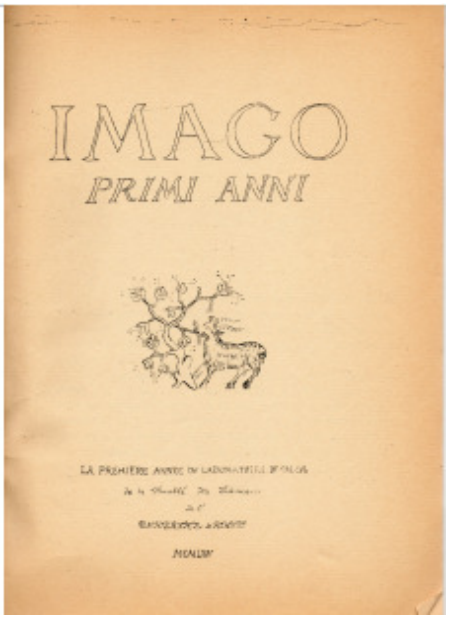
Birth of CA in Rennes. What's New 40 Years Later?

Brigitte Le Roux

MAP5/CNRS, Université Paris Descartes & CEVIPOF/CNRS, Sciences-Po Paris
Brigitte.LeRoux@mi.parisdescartes.fr

In this talk, I will give an overall picture of the birth of Correspondence Analysis (CA) in Rennes, by relying on the 1964-65 mimeographed reports published at the "Centre de Calcul Automatique de la Faculté des Sciences de Rennes" (Computer Center).

The origins of CA date back to 1940 and 1941 at least when Guttman and Fisher established the statistical characterizations of the method. But it was J-P. Benzécri who, during the sixties, made it the pioneering method of the French approach to Data Analysis, with the *geometric presentation* and the *aids to interpretation* and who gave it its definitive name: "*analyse des correspondances*".



Although I will not examine all the texts of that period, I will show examples of the first analyses made with the computing means that were then at our disposal: an IBM 1620 computer. I will add some theoretical texts to this brief inventory so as to show the emergence of the *three key ideas* of "l'analyse des données", that are geometric, formal and descriptive.

Following the Rennes period a complete methodology was developed around Correspondence Analysis. It placed the emphasis on the representation of a table of numbers by clouds of points in a multidimensional geometric space. This is that we — Henry Rouanet and I —called Geometric Data Analysis (GDA) following the suggestion of Patrick Suppes (Stanford University).

Since the very beginning, GDA has been applied to a large range of domains, such as medicine, lexicology, marketing research, econometrics and social sciences. In the latter domain, the work of Pierre Bourdieu is exemplary in regard to the "elective affinities" between the spatial conception of the social space and GDA representations. These affinities led Bourdieu and his school to use CA (especially MCA) consistently since 1976.

In conclusion, I will evoke recent developments of GDA, especially for MCA, in connection with research in social sciences

References

Imago Primi anni (1964), Cours de Linguistique (1964), Recueil Psychologique (1964), Sur l'Instauration d'un Code (1965), *mimeographed reports,* Faculté des Sciences de Rennes.

BENZÉCRI J-P. (1982). *Histoire et Préhistoire de l'Analyse des Données*, Dunod, Paris.

CORDIER-ESCOFIER B. (1964), L'analyse factorielle des correspondances, *PhD Dissertation*, Faculté des Sciences de Rennes, published in *Cahiers du BURO, 13 '1969).*

MURTAGH F. (2005). *Correspondence Analysis and Data Coding with Java and R*, Chapman & Hall, London (Foreword by J-P. Benzécri).

LE ROUX, B.& ROUANET, H. (2004). *Geometric Data Analysis : From Correspondence Analysis to Structured Data Analysis*, Kluwer, Dordrecht (Foreword by P. Suppes).

LE ROUX, B. & ROUANET, H. (2010). *Multiple Correspondence Analysis*, QASS volume 163, SAGE, CA : Thousand Oaks.

Unifying the geometry of simple and multiple correspondence analysis

Michael Greenacre[1,*]

1. Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, Barcelona, 08005 SPAIN
* michael@upf.es

Keywords: Multiple correspondence analysis, optimal scaling, adjusted inertias, contributions, joint correspondence analysis.

There are two different approaches to the definition and interpretation of multiple correspondence analysis (MCA): the first can be called the *scaling approach*, following Guttman's earlier work and manifest principally in the Gifi system (see, for example, Michailidis & de Leeuw, 1998), and the second the *geometric approach*, as promoted principally by Benzécri and his followers. While the scaling approach generalizes simply from the bivariate to the multivariate context, the popular geometric approach is fraught with inconsistencies, a topic that has already generated quite a lot of discussion.

In this talk I discuss alternative ways of generalizing the geometry of simple correspondence analysis (CA) to MCA. The Burt matrix is the key concept here, and there are mainly two possibilities: joint correspondence analysis (JCA) and what I call adjusted MCA (for a discussion of both of these, see Greenacre, 2007: chapters 18–20). Both have simple CA as special cases when the number of categorical variables is two. While each of these alternatives has its advantages, unfortunately neither is perfect in all its characteristics. Having to choose, my preference would be for adjusted MCA, since it preserves the optimal scaling properties of the solution, while coming as close as possible to the JCA solution which optimally fits all two-way cross-tables. This compromise solution is thus the default option provided in our **ca** package in R (Nenadić & Greenacre, 2007).

Apart from explicitly defining these two approaches, I will define (and illustrate with an application) (i) how percentages of explained variance (i.e., inertia) are calculated in each case, (ii) how to scale the solutions, (iii) how to compute contributions of each point to the dimensions and of each dimension to the points, and (iv) how supplementary category points are displayed. All of these computational aspects, some of which are new, are included in the latest version of the **ca** package, released at CARME 2011.

References

Greenacre, M. (2007). *Correspondence Analysis in Practice, Second Edition*. Chapman & Hall / CRC Press, London.

Michailidis, G. & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, **13**, 307–336. Available for download at http://projecteuclid.org

Nenadić, O. & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: the **ca** package. *Journal of Statistical Software*, **20(3)**.
URL http://www.jstatsoft.org/v20/i03/

Graph Partitioning by Correspondence Analysis and Taxicab Correspondence Analysis

Choulakian Vartan[1,*], De Tibeiro Jules[1]

1. Université de Moncton, Moncton, NB Canada
* Contact author: vartan.choulakian@umoncton.ca

Keywords: Network analysis, Graph partitioning, Graph Laplacian matrix, Correspondence analysis, Taxicab correspondence analysis

We consider correspondence analysis (CA) and taxicab correspondence analysis (TCA) of relational datasets that can mathematically be described as weighted loopless graphs. Such data appear in particular in network analysis, see for instance Kolaczyk (2009). Benzecri (1973, chapters 9 and 10) discuss CA of such data sets, where the influence of the diagonal elements on the factors and dispersion measures is emphasized and quantified. We present CA and TCA as relaxation methods for the graph partioning problem as described in Ding (2004) and von Luxburg (2007). Examples of real datasets are provided.

References

Benzecri J.P. (1973). L'Analyse des Données: Vol. 2: L'Analyse des Correspondances. Dunod, Paris.
Choulakian V. (2006). Taxicab correspondence analysis. *Psychometrika* **71**: 333-345.
Choulakian V. (2008). Multiple taxicab correspondence analysis. *Advances in data Analysis and Classification*, **2**, 177-206.
Ding C. (2004). *A tutorial on spectral clustering. Talk presented at ICML.* http://crd.lbl.gov/~cding/Spectral/
Kolaczyk E.D. (2009). Statistical Analysis of Network Data. Springer: N.Y.
Lebart L. (2001). Classification et analyse de contiguité. *La Revue de Modulad* **27** :1-22.
von Luxburg U. (2007). A tutorial on spectral clustering. *Statistics and Computing* **17**, 395-416.

Singly and Doubly Ordered Cumulative Correspondence Analysis.

L. D'Ambra[1,*] **, E. J. Beh**[2]**, I. Camminatiello**[1]
1. University of Naples Federico II
2. University of Newcasle
* Contact author: dambra@unina.it

Keywords: Taguchi's statistic, doubly cumulative chi-squared statistic, correspondence analysis

The classical approach to correspondence analysis (CA) is designed to allow its user to a graphically summarize the association between two or more categorical variables that form a contingency table. Despite its popularity and utility, the classical approach does not take in consideration the structure of ordered variables. One way to performing CA when the variables have an ordered structure is to consider the Taguchi's statistic (Taguchi, 1974). Beh, D'Ambra, Simonetti (2010) demonstrated the applicability of considering this statistic which takes into account the ordered structure by considering the cumulative sum of cell frequencies across the variable. Thus, the statistic is defined by summing the chi-squared statistic for each $I \times 2$ contingency table obtained by aggregating the column categories *1* to j and aggregating the column categories *(j+1)* to J. For this reason, the Taguchi's statistic is also referred to as cumulative chi-squared statistic (Nair; 1987).

Cuadras (2002) proposes an approach to correspondence analysis based on double cumulative frequencies. However, it does not decompose any known index. In this paper we explore a generalization of Taguchi's statistic which takes into account the presence of two ordinal categorical variables by considering their cumulative sum of cell frequencies. This generalization is analogous to the doubly cumulative chi-squared statistic which is constructed by summing the chi-squared statistic for each 2×2 sub-table formed by pooling adjacent rows and columns of the original contingency table; see Hirotsu (1986).

We illustrate this approach to CA using a partition of the statistic proposed by Hirotsu. Its application presents some interesting properties and allows the analyst to represent the variations of row and column categories rather than the categories on the space generated by cumulative frequencies.

References

Beh, E. J., DAmbra, L. & Simonetti, B. (2010). Correspondence analysis of cumulative frequencies using a decomposition of Taguchis statistic. *Communications in Statistics Theory and Methods* (to appear).

Cuadras, C. M. (2002). Correspondence analysis and diagonal expansions in terms of distribution functions. *J. of Statistical Planning and Inference*, **103**, 137–150.

Hirotsu C. (1986). Cumulative Chi-squared Statistic as a Tool for Testing Goodness of Fit. *Biometrika*, **73**, 165–173.

Nair, V. N. (1987). Chi-squared type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association*, **82**, 283–291.

Taguchi, G. (1974). A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. *Saishin Igaku*, **29**, 806–813.

Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis

Julie Josse[1*], **Marie Chavent**[2], **Benoît Liquet**[3], **François Husson**[1]
1. Agrocampus, 65 rue de St-Brieuc, 35042 Rennes, France
2. Université V. Segalen Bordeaux 2, 146 rue L. Saignat, 33076 Bordeaux, France
3. Equipe Biostatistique de l'U897 INSERM, ISPED
* Contact author: josse@agorcampus-ouest.fr

Keywords: Multiple Correspondence Analysis, Categorical Data, Missing Values, Imputation, Regularization

A common approach to deal with missing values in Exploratory Data Analysis consists in minimizing the loss function over all non-missing elements. This can be achieved by EM-type algorithms where an iterative imputation of the missing values is performed during the estimation of the axes and components. This presentation proposes such an algorithm, named iterative MCA, to handle missing values in Multiple Correspondence Analysis (MCA). This algorithm, based on an iterative PCA algorithm, is described and its properties are studied. We point out the overfitting problem and propose a regularized version of the algorithm to overcome this major issue. Performances of the *regularized iterative MCA* algorithm are assessed from both simulations and a real dataset. Results are promising for MAR and MCAR values (Little and Rubin, 1987, 2002) with respect to other methods such as missing-data passive modified margin, an adaptation of missing passive method used in Gifi's Homogeneity analysis framework.

References

M. Greenacre & R. Pardo (2006). Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological methods and research*, 35 (2):193–218.

J. Josse, J. Pagès & F. Husson (2009). Gestion des données manquantes en analyse en composantes principales. *Journal de la Société Française de Statistique*, 150: 28–51.

R. J. A. Little & D. B. Rubin (2002). *Statistical Analysis with Missing Data.* Wiley series in probability and statistics, New-York.

J. Meulman (1982). *Homgeneity Analysis of Incomplete Data.* D.S.W.O.-Press, Leiden.

Y. Takane & H. Hwang (2006). *Regularized multiple correspondence analysis.* In J Blasius and M J Greenacre, editors, Multiple Correspondence Analysis and Related Methods, pages 259–279. *Chapman Hall.*

P.G.M. van der Heijen & B. Escofier (2003). *Multiple correspondence analysis with missing data.* In Analyse des correspondances. *Presse universitaire de Rennes.*

Correspondence analysis and moderate outliers

Anna Langovaya[1,*] **, Sonja Kuhnt**[1]
1. TU Dortmund University, Faculty of Statistics
* Contact author: langovaya@statistik.tu-dortmund.de

Keywords: Correspondence analysis, outliers, multi-way contingency tables.

The Correspondence Analysis (CA) is a popular method for analysis of categorical data. In CA as well as in every statistical analysis, observations can appear that seem to deviate strongly from the majority of the data. Such observations are usually called outliers and may contain important information about unknown irregularities, dependencies and interactions within the data. However, behavior of CA in the presence of outliers in the table is not sufficiently explored in the literature, especially in the case of multidimensional contingency tables.

We will be studying more subtle cases of outliers, which are not immediately suspicious in the table based on their size, but play a crucial role for the statistical analysis. We apply CA (Benzecri (1992), Blasius and Greenacre (2006)) to three-way contingency tables with dependent entries, where specific dependencies are caused by outliers of moderate size. In our work outliers are chosen in such way, that they break independence in the table, but cannot be spotted immediately.

We study the change in the CA row and column coordinates caused by one or more outliers. We also perform numerical analysis of CA coordinates and suggest possible criteria for identifying hidden outliers in multi-way contingency tables.

References

Benzecri, J.-P. (1992). Correspondence analysis handbook. *Marcel Dekker, Inc., New york*.

Blasius, J. and Greenacre, M. (2006). Multiple Correspondence Analysis and Related Methods. *Chapman & Hall, London*.

A Comparison between Latent Semantic Analysis and Correspondence Analysis

Julie Séguéla[1,2] **, Gilbert Saporta**[1,*]
1. CEDRIC, CNAM, 292 rue Saint-Martin, F-75141 Paris cedex 03
2. Multiposting.fr, 33 rue Réaumur, 75003 Paris
* Contact author: gilbert.saporta@cnam.fr

Keywords: Latent semantic analysis, textual data, correspondence analysis, web data

Latent Semantic Analysis (LSA) is a technique for analyzing textual data through a singular value decomposition of term-document matrices (Deerwester et al. (1990), Landauer et al. (2007)). The basic postulate is that there is an underlying latent semantic structure in word usage data that is partially hidden or obscured by the variability of word choice (synonymy problem). LSA is also called Latent Semantic Indexing (LSI) in information retrieval, where the main application consists in computing similarities between user's query and all documents in the space, or between documents.

Since LSA is a SVD of a contingency table, it strongly resembles to Correspondence Analysis (CA), see Lebart et al. (1998). Before performing the SVD, practitioners of LSA recommend several weighting functions of the frequencies, but not the one leading to the chi-square metric. Typically, LSA allows to reduce the dimensionality from several thousands to several hundred of a huge but sparse data matrix. Given the dimension, graphical representations are useless. In the context of statistical implementations, the coordinates can be used for categorization tasks (in supervised or unsupervised frameworks).

We first compare basic LSA with CA on a toy example. Then performances of CA and LSA with several weighting functions are compared on a large data set coming from job offers posted on the web. When posted on the internet, job offers have been labeled by recruiters according to the job category (e.g. Marketing, Information Systems, Finance, etc.). We are interested in the capacity of these document representation technics to lead us to the real job category with a clustering method. After preprocessing of job offers, we compute similarities between texts based on coordinates in reduced spaces and apply an hybrid method combining hierarchical clustering and k-means algorithm. Performance of text representation methods will be assessed with three different measures (Cohen's Kappa, Rand index, F-measure) and discussed according to the number of dimensions kept.

References

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science*, **41**, 391-407.

Landauer, T. K., & al. (2007). *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates.

Lebart, L., Salem, A. & Berry, L. (1998). *Exploring Textual Data*, Kluwer.

LSA website,
`http://lsa.colorado.edu/`.

Multiblock method for categorical variables. Application to air quality in pig farms.

Stéphanie Bougeard[1,*], **El Mostafa Qannari**[2], **Christelle Fablet**[1]
1. French agency for food, environmental, and occupational health safety (Anses), Department of Epidemiology - Zoopole, BP53, 22440 Ploufragan, France
2. Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering (Oniris), Department of Chemometrics and Sensometrics - Rue de la Géraudière BP 82225, 44322 Nantes Cedex, France
* Contact author: stephanie.bougeard@anses.fr

Keywords: Categorical discriminant analysis, Multiblock Redundancy Analysis, Multiblock Partial Least Square, Multiple Non-Symmetrical Correspondence Analysis, *Disqual* procedure.

Research in veterinary epidemiology is often concerned with predicting a categorical variable related to animal health, from a large number of categorical variables (*i.e.* the potential risk factors for the disease under study) related to the breeding environment, alimentary factors and farm management, amongst others. In a more formal way, the aim of the study is to explain a categorical variable y by a large number of K other categorical explanatory variables $(x_1, \ldots, x_K)$, all these variables being measured on the same statistical unit. In veterinary epidemiology, the statistical procedures usually performed are particular cases of Generalized Linear Models, especially logistic models. But all the potential explanatory variables, in addition to being redundant, cannot be included in a single model. Considering the aim and the specificity of veterinary data, our research work focuses on methods related to the multiblock modelling framework, each block being formed of the indicator matrix associated with each categorical variable. The well-known conceptual models are the Structural Equation Modelling (SEM) and the PLS Path Modelling, which have been recently extended to categorical data. For our purpose of exploring and modelling the relationships between one categorical variable with several other ones, simpler procedures can be used, such as multiblock $(K + 1)$ methods. Multiblock Partial Least Squares (Wold, 1984) is a widely-used multiblock modelling technique. It is not originally designed as a discrimination tool, but it is used routinely for this purpose in the two-block case. A categorical extension of multiblock Redundancy Analysis, as an alternative to multiblock PLS, is proposed (Bougeard, In Press). The main idea is that each indicator matrix is summed up with a latent variable which represents an optimal coding of the associated categorical variable. This can be related to the measurement model of SEM, which relates observed indicators to latent variables. In addition, a structural model is built, which specifies the relations among latent variables. All the latent variables, from measurement and structural models, come from a well-identified global optimization criterion which leads to an eigensolution. A comparison of the categorical multiblock Redundancy Analysis with the main alternative methods is undertaken. In practice, this method mainly competes with methods belonging to the class of Generalized Linear Models (*e.g.* logistic and PLS logistic regression) and other methods that can be viewed as categorical extension of multiblock methods (*e.g.* categorical extension of multiblock PLS, the *Disqual* procedure (Saporta, 2006), the Multiple Non-Symmetrical Correspondence Analysis). Practical uses of the proposed method are illustrated using an empirical example in the field of veterinary epidemiology. The aim is to study the air quality in pig farms (coded in three categories: cold, temperate, temperate with gases) in the light of nineteen potential explanatory categorical variables, related to the heating and the ventilation systems, the management practices and the farm structure. Risk factors for inappropriate air quality are given. It is concluded that categorical multiblock Redundancy Analysis is a relevant tool for qualitative discrimination. Moreover, all the interpretation tools developed in the multiblock framework can be adapted to enhance the interpretation of categorical data and unveil new information for the user. The multiblock methods can be directly adapted to more complex data, thus extending the strategy of analysis to the prediction of several categorical variables.

References

Bougeard, S., Qannari, E.M., Lupo, C. & Hanafi, M. (In Press). From multiblock partial least squares to multiblock redundancy analysis. A continuum approach. *Informatica*.

Saporta, G. & Niang, N. (2006). Correspondence analysis and classification. In: *Multiple correspondence analysis and related methods*. Greenacre, M. & Blasius, J. Eds, Chapman & Hall, pp. 372–392.

Wold, S. (1984). Three PLS algorithms according to SW. *MULDAST*, Umea University, Sweden, pp. 26-30.

Towards the integration of biological knowledge with canonical correspondence analysis when analyzing Xomic data in an exploratory framework

Marie Verbanck[1,*] **, Sébastien Lê**[1] **, Jérôme Pagès**[1]
1. Agrocampus Ouest, Laboratoire de Mathmatiques Appliques, Rennes, France
* Contact author: mverbanc@agrocampus-ouest.fr

Keywords: transcriptomic data, integration of biological knowledge, canonical correspondence analysis, multiple factor analysis

Post-genomic data present a strong character of exhaustiveness, as the microarray technology allows to monitor the expression of potentially all the genes within a tissue. All the gene expressions are measured without a priori, regardless of any biological hypothesis on the gene's behavior according to the experimental conditions of interest. When analyzing data with a characteristic of exhaustiveness, multivariate exploratory analysis, such as principal component analysis (PCA), establishes itself to consider simultaneously the whole information.

The collection of transcriptomic profiles permits to focus on the variability, among individuals, described by the expression of their genes. This leads to particular datasets, as the number of individuals is highly superior to the number of variables. As a matter of fact, the correlation circle appears to be extremely encumbered.

Therefore, we propose a two-step approach to analyze those kinds of data. Firstly, the entire dataset is taken into account without any statistical or biological selection, thanks to a multivariate exploratory analysis. Secondly, we propose to use external biological knowledge, into supplementary elements, to facilitate the interpretation of the correlation circle. The biological knowledge, in the form of Gene Ontology terms, associates to a gene its biological functions. Thus, the biological knowledge is used to build modules of genes, which are projected as supplementary elements: the main dimensions of variability are no longer interpreted at a gene level but rather at a modular level.

In this talk, we intend, to present a framework for the interpretation of multivariate exploratory analysis results, such as PCA, when applied to Xomic data. Then we propose to explore several approaches which make use of biological knowledge to constitute modules of genes. We will particularly focus on canonical correspondence analysis (CCA), as it appears to be particularly adapted to build modules of genes implicated into the same biological processes, under condition of co-expression.

Consequently, CCA is used here to define a new distance between the genes: two genes are closed if they are involved into the same biological processes, upon condition that they are co-expressed into the experiment. Then hierarchical classification is used to obtain groups of genes, which will be projected as supplementary elements.

The interpretation of the biological processes is thus facilitated by the co-expression of the genes within a group, whereas the method highlights a few key-genes whose functions can be easily taken into account to go deeper into the interpretation. An application of this method to a chicken microarray data set has allowed to bring out the well-known mechanisms implemented in reply to fasting, and to come up with new trails.

References

De Tayrac, M., Lê, S., Aubry, M., Mosser, J., Husson, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, **2009**, 10–32.

Ter Braak, Cajo J. F. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**, 1167–1179.

Escofier, B., Pagès, J. (1990). Analyses factorielles simples et multiples, objectif, méthodes et interprétation. *Paris, Dunod*.

Betweenness relation orientated by Guttman effect in critical edition

Marc Le Pouliquen[1] **, Marc Csernel**[2]
1. Telecom Bretagne, Labsticc UMR 3192 , BP 832, 29285 Brest Cedex - France
2. Inria-Rocqencourt, BP-105- 78180 Le Chesnay - France
* Contact author: marc.lepouliquen@telecom-bretagne.eu - marc.csernel@inria.fr

Keywords: Betweenness, Guttman, Critical Edition, Filiation of manuscripts, Seriation,

The goal of this paper is to model the ternary betweenness relation within the framework of critical edition of manuscripts. The editor tries to reconstruct, as well as possible, the original manuscript using a corpus of various preserved manuscripts. This corpus is made up from manuscripts which have been copied one from the other. To achieve such a goal, it is interesting to draw up a family of filiations trees called "stemma codicum". As suggested by Don Quentin, we propose to build this tree using the betweenness relation within the manuscripts. Manuscript B is between manuscripts A and C, if manuscript C was copied from manuscript B which itself was copied from A. We notice that the number of betweenness relations grows rather quickly with the number of manuscripts to be compared. It is usually too large to allow hand made comparison and construction. Thanks to the calculation capabilities of current computers, the method of Don Quentin can be modified and adapted to build the stemma by computers. We finaly observe that these relations provide a seriation of the manuscripts set which can direct an editor towards a text which is rather close to the original one. To acquire the seriation from betweenness relations, we use Guttman effect to choose preponderant relations among all.

References

Benzécri J.-P. & coll. (1973) - La taxinomie, Vol. I ; L'analyse des correspondances, Vol. II, *Dunod*, Paris.

Lerman I. C. (1972) Analyse du phénomène de la sériation à partir d'un tableau d'incidence, *Math.Sci. Humaines*, **38**, 39-57.

Menger K. (1928) Untersuchungen ber allgemeine Metrick, *Mathematische Annalen*, **100**, 75-163.

Restle F. (1959), A metric and an ordering on sets, *Psychometrika*, **24**, 207-220.

Quentin H. (1926). Essais de critique textuelle, *Picard*.

Weisz Communication Styles Inventory (WCSI-Version 1.0): Development and Validation

Robert Weisz[1] , Jahanvash Karim[2] *

1. CERGAM, IAE d'Aix en Provence, Université de Paul Cézane. Clos Guiot Puyricard – BP 30063, Aix-en-Provence Cedex 2 , France.
2. Doctoral student, CERGAM, IAE d'Aix en Provence, Université de Paul Cézane, France.
* Contact author: j_vashl@hotmail.com

Keywords: Communication Styles, Child Development, Inventory, Validation

We all communicate differently and have different needs in meeting. Managers often find it helpful to have a model for recognizing different communication styles of people within organizational settings. By having an insight into the communication styles of employees, as well as their predominant needs, a manager may be better able to detect and resolve dysfunctional behaviours. Unfortunately, there has been a lack of a psychometrically sound yet practically short communication styles measure for management research. The purpose of this study was to develop such a measure and provide evidence concerning its validity.

Borrowed heavily from child development theories , we proposed that (a) there are certain set of needs which children express and desire to satisfy during different stages of their development-affection during infancy, attention during toddlerhood, structure and limits during preschool, and esteem during middle childhood; (b) the same set of needs, one way or other, predominantly guide human behavior during adulthood; (c) people from their childhood relational experiences and satisfaction and non-satisfaction of these *needs*, develop cognitive representations, or internal working models, that consist of specific ways of expressing and satisfying these needs; (d) people mainly use four psychological languages or communication styles for satisfying their different set of needs, that is, relationships (R) for affection, ideas (I) for attention, structures (S) for confirmation, and values (V) for esteem; and finally (e) the frequent use of any particular psychological language or communication style depends on the importance of certain set of needs for an individual.

Construction of the communication styles inventory (CSI) occurred in three major phases. In the first phase of constructing the scale, we generated a pool of 152 short phrases and adjectives organized in 38 frames of four choices each. Each choice within each frame reflected an adaptive tendency towards a particular communication style (i.e., R, I, S, or V). Respondents selected a forced choice option of "most-like me" (one choice among the four). To explore the inherent structure of the 38-item scale the Multiple Correspondence Analysis-MCA method was applied to the response data set of N= 1453. Initial visualization of joint plot of category points and discrimination indices revealed that 23 items performed poorly in discriminating among the item options or styles. Thus, these 23 items were dropped from further analysis and we continued with a set of remaining 15 items. The Cronbach alph, based on optimal scaling technique, revealed to be .81 for the set of 15 items. In the second step, we subjected the response patterns on the items to latent class cluster analysis (LCA). The major goal of LCA is to determine the number of latent classes *R*- in this case, communication styles- that are necessary to account for the association that exists among the manifest variables. Theoretically, if our 15-items scale discriminates well among the four communication styles (R, I, S, V), we might expect to see a four cluster solution. Latent class models were tested for 1 to 6 groups of latent classes. LCA results clearly supported a four class solution representing four communication styles.

To establish the construct validity of the communication styles inventory, another study was conducted to test the relationships between scores on communication styles inventory with other established constructs, that is, the Big Five personality dimensions and emotional intelligence. Participants of this study included 228 students from two nonnative English speaking national cultures: 101 from a university in Aix-en-Provence, France (45 males, 56 females), and 127 from a large university in the province of Balochistan, Pakistan (78 males and 48 females, one unreported). Results indicated that 15-items communication styles inventory is related to but yet different from the Big Five personality dimensions and emotional intelligence.

Application of Correspondence Analysis and Related Methods in Evaluation of Knowledge and Skills of Young People

Prof. Jozef Dziechciarz[1] , dr Alicja Grześkowiak[1], dr Agnieszka Stanimir[1]

1. Wroclaw University of Economics, Poland
* jozef.dziechciarz@ue.wroc.pl, alicja.grzeskowiak@ue.wroc.pl, agnieszka.stanimir@ue.wroc.pl

Keywords: multiway correspondence analysis, multidimensional statistical analysis, knowledge and skills of young people

The analysis of knowledge and skills of a young person is an extremely important task in the educational process. The direct effect is the possibility to support the creation and orientation of educational and professional development paths. It is also crucial to properly identify relationships between the level of knowledge and various aspects of life (demographic, social, economic, etc.) from the perspective of both authorities constituting educational policy and teachers. With comprehensive information, teachers are able to help young people with the choice concerning further education and students can gain reliable information about their perspectives. This paper attempts to analyze the level of knowledge and skills of young people at regional level (Lower Silesia) and global level (Europe).

Variables describing the skills and competences as well as socio-economic factors are often nominal or ordinal. It is therefore natural to apply correspondence analysis and related techniques to identify associations between categories or relationships between variables.

References

Blasius, J. (2001). *Korrespondenzanalyse*. Munchen: Oldenbourg Verlag.

Education at a glance. OECD indicators. (2009): OECD Report

Greenacre, M., J. (1984). *Theory and applications of correspondence analysis*, London : Academic Press.

Information Sources that EU Tourists Use:

A Cross-country Study

Tor Korneliussen

Bodø Graduate School of Business, 8049 Bodø, Norway

Tor.Korneliussen@hibo.no

Key words: Information search, European Union, cross-country analysis, marketing research

The ability to attract tourists is crucial for the financial success of travel destinations. Especially in marketing research there is much interest in which information sources tourists use when selecting a destination (Gursoy and Chen, 2000; Gursoy and Umbreit, 2004). The purpose of this study is to investigate tourists' use of information sources when making decisions about their travel/holiday plans and to try to shed light on to what degree and why tourists from different countries have varying information source behaviour. The emphasis of the paper is on information search behaviour among tourists from the 27 member countries of the EU, the total sample size is n=27.000. This study investigates which information sources European tourists use when making decisions about their travel/holiday plans.

The analysis starts by applying simple correspondence analysis to information sources and countries, it proceeds by applying simple correspondence analysis to individual level variables such as gender, age, education and occupation by information sources. To include the interactions effects the study turns to multiple correspondence analysis and shows the relationships between individual level data and information sources, using countries as supplementary variables. Several possibilities to analyze this kind of data with the help of correspondence analysis will be discussed.

References

Gursoy, D. and Chen, J.S. (2000). Competitive analysis of cross-cultural information source behavior. *Tourism Management*, 21(6), 583-590.

Gursoy, D. and Umbreit, W.T. (2004). Tourist information source behaviour: cross-cultural comparison of European Union member states, *International Journal of Hospitality Management*, 23(1), 55-70.

About the history of Multiple Correspondence Analysis

Ludovic Lebart
Telecom-ParisTech
www.lebart.org
lebart@telecom-paristech.fr

The CARME 2011 venue (Rennes) has witnessed in 1965 the dissertation of the late Brigitte Cordier-Escofier (Sur l'analyse des correspondances; directed by Jean-Paul Benzécri). It provides an opportunity to call to mind the upsurge of exploratory multivariate data analysis that follows the year 1965 (under the name: "Analyse des données" in francophone countries). Multiple Correspondence Analysis (MCA) has emerged as such in the seventies, but its history, indissoluble from the history of Correspondence Analysis (CA), of which it is a variant, goes back much further in the past. The history of MCA and the simultaneous revival of induction in statistics will then be a pretext to talk about the giants on whose shoulders we were seated...

Some Comments on Correspondence Analysis

Pierre Cazes
CEREMADE, Université Paris Dauphine
* Contact author: cazes@ceremade.dauphine.fr

After having recalled why correspondence analysis (CA) and more generally Data Analysis can be considered as an experimental science, we will analyze the activity of the Laboratory of Statistics of Professor Benzécri at University Pierre et Marie Curie (Paris 6) in the seventies and the eighties, and in particular the Master of Statistics and the publications that have been released. We will then come back to the importance of coding in CA, and especially fuzzy coding and coding that allows obtaining the equivalence between CA and other analyses. Then, we recall that CA is a particular case of numerous classical analyses, and we will detail the case of multiple tables. We will speak about the link between ascending hierarchical classification and CA. Then, we will analyze the links between CA and classical statistics. We show interest of CA in certain modeling problems and treat briefly the use of CA in the working environment. We won't try to be exhaustive in this presentation. We just highlight some important points on CA without quoting all the possible references on a given subject.

Cluster analysis with k-means: what about the details ?

Maurice Roux
Université Paul Cézanne
Marseille, France

Background:
When using the k-means procedure (and its variants) there are several parameters to select beforehand, the main one being the number of clusters. The usual strategy to to determine this number is to repeat the whole procedure with various cluster numbers and to select the one which leads to the best fit between the resulting partition and the initial data.

To evaluate this fit a number of indexes (internal criteria) have been proposed in the literature. In addition, for a fixed number of clusters it is recommended to restart "many" times the overall computations with new random initializations.

The present paper, based on both artificial and real life data, wants to help for the choice of a goodness-of-fit index and put forward some guidelines for the number of restarts.

Main results :
Three indexes do give consistent appreciations, namely Dunn's index, Kendall's tau and the contingency Khi-square based on the quadruples (pairs of pairs of objects). As for the second target parameter, it appears that the number of restarts is not a key parameter, since the "best" results are quickly reached after, say, a few tens of repeated random initial partitions.

Incidentally, after a multiple restart k-means it is very useful to run a correspondence analysis program applied to a consensus matrix over the objects. Such an analysis clearly detects those objects not included in any cluster which may be tagged as "unclassifiable". More over it confirms or invalidates the number of clusters.

When there exists a known partition of the data it may be tempting to use it as a reference to evaluate indexes and clustering methods. But an example in gene expression data shows this approach is questionable.

Conclusion:
The k-means clustering process is a very useful method, able to deal with very big data sets. It is even more efficient when a good quality index is used to establish the number of clusters. The present work is not really a benchmark but it emphasizes the difficulty of finding groups in real life data sets. The use of correspondence analysis with a consensus matrix greatly helps to discover "unclassifiable" observations which often confuse the clustering results.

The Chemical Analysis of Soil – Plant With High Boron Concentrations by Log-Ratio Analysis

Harun BÖCÜK[1] Zerrin AŞAN[2*] Cengiz TÜRE[3]

1. Anadolu University, Science Faculty,Biology Department,Turkey
2. Anadolu University, Science Faculty,Statistic Department,Turkey
3. Anadolu University, Science Faculty,Biology Department,Turkey
*Contact author: zasan@anadolu.edu.tr

Keywords: Boron concentration, composotional dat**a**, log-ratio analysis

Compositional data consisting of vectors of positive components subject to a unit-sum constraint arise in many disciplines, for example, in geology as major oxide compositions of rocks, in sociology and psychology as time budgets, that is parts of a time period allocated to various activities, in politics as proportions of the electorate voting for different political parties, and in genetics as frequencies of genetic groups within populations (Aitchison, 1994). Log-ratio analysis applies to any table of strictly positive data, where all data entries are measured on the same scale (Greenacre, 2010).
In this study 7 boron reserve areas were investigated in Turkey. Although extractable boron range is a limiting factor for many plant species, 10 plant taxa which can distribute onto the soils with over extractable boron level were determined. Boron accumulation and germination characteristics of these taxa in different boron levels were also studied. Boron concentration together with N, P, K, Na, Ca and Mg proportions in the soil at the sample areas were determined by chemical analysis. In addition, the same process was repeated for plants that grew around boron reserves (Böcük, 2010). In this study log-ratio analysis was applied for chemical analysis of soil-olant with high boron concentrations.

References

Aitchison, J. (1994). Principles of compositional data analysis. *Multivariate Analysis and its Applications*, **24**, 73-81**.**

Böcük, H. (2010*). Investigation of Natural Plant Diversity on the Soils With High Boron Concentrations in Terms of Soil-Plant Relations in West Anatolia.* Doctoral Thesis Anadolu University, Eskişehir, Turkey.

Greenacre, M. (2010). *Biplots in Practice*, BBVA Foundation.

Evaluation of Turkish Media and The Athletics News by Correspondence Analysis

Hasan Durucasu[1,*] **, Özgür İcan**[2]
1. Professor, Anadolu University, F.E.A.S, Dept. of Business Administration
2. Res. Assistant, Anadolu University, F.E.A.S, Dept. of Business Administration
* Contact author: hdurucasu@anadolu.edu.tr

Keywords: Athletics News, Correspondence Analysis

In Turkey, the general tendency towards the sports news usually involves football (soccer). Much of the attention is given to players and coach performances, even the players? private lives occupy a huge amount of space in a newspaper. The athletics news usually stays in the background. Most of the athletics news is not in the main sports pages but on some other pages with small amount of spaces and overlooking most details.

In this study, the Turkish newspapers are reviewed for athletics news. Initially general structure of the news itself and the amount of information given about an athletics organization is investigated by classical Unix tools (i.e. Bash scripts containing heavy usage of `grep`, `sed`, `awk` and alike). Secondly, the categories of athletics news published according to months is obtained by content analysis. Finally it is determined relationship between these categories and months by correspondence analysis.

References

Daddario G. (1994). Chilly Scenes of the 1992 Winter Games: The Mass Media and the Marginalization of Female Athletes. *Sociology of Sport Journal*, **11,3**, 275–288.

Gusfield, J.R. (2000). Sport as Story: Form and Content in Athletics. *Culture and Society*, **37**, 63–70.

Gardnier, G. (2003). Australian Print Media Representation of Indigenous Athletes in the 27th Olympiad. *Journal of Sport & Social Issues*, **27,3**, 233–260.

The insurability of risks: A quantitative approach applied to motor insurance

Kmar Fersi[1,*], **Noureddine Benlagha**[2] **and Samir Ben Ammou**[3]
1. Institut Supérieur de Gestion Sousse , Faculté des Sciences, Computational Mathematics Laboratory, Route de Kairouan, 5019 Monastir, Tunisia
2. IHEC Sfax-Tunisia, Université Paris2, ERMES-UMR7181-CNRS, 12 place du Panthon, 75005 Paris, France
3. Faculté des Sciences, Computational Mathematics Laboratory, Route de Kairouan, 5019 Monastir, Tunisia
* Contact author: fersi.gmar@yahoo.fr

Keywords: Extreme Value Theory, Car Insurance, Premium, Risks Insurability

This work examines the limits of insurability and their implications. Using quantitative approaches, we show how insurers can overcome obstacles and develop new types of risk coverage best suited to customer needs.
These approaches allow us to quantify the behavior and impact of extreme losses in a portfolio of an insurance companies. Insurers are used to calculate an estimate of the pure premium based on classical methods. In this study we implement a new method to calculate a pure premium which is more adequate to cover extreme losses that exceed a certain threshold. This value can be considered, first, as a maximum premium accepted by the insured, and second, as the appropriate premium to face the risk of ruin of the insurer.
This methodology has been applied to a real non life problem; data from French automobile insurance. We also propose a pricing strategy to insurers.

References

Beirlant, J., Goegebeur, Y., Segers J., and Teugels, J. (2004). Statistics of Extremes: Theory and Applications. Wiley.

Berliner, B.(1982). Limits of insurability of risks. Prentice Hall, Swiss Re Economic Research and Consulting.

Davison, A., and Smith, R. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B.* **52**, p. 393-442

Denuit, M., and Charpentier, A. (2004). Mathématiques de l'Assurance Non-Vie. Tome I: Principes Fondamentaux de Théorie du Risque. Collection Economie et Statistique Avancées, Economica, Paris.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications.* **75**, p. 149-172.

Embrechts, P., Claudia, K. and Thomas, M. (1997). Modelling Extremal Events For Insurance and Finance. New York: Springer. 1st ed.

Falk, M., and al. (2004). Lows of Small Numbers: Extremes and Rares Events. 2nd ed: Birkhäuser.

Fromont, E. (2005). *Modélisation des rentabilités extrêmes des distributions de Hedge Funds.* CREM UMR CNRS 6211-Axe Macroéconomie et Finance.

Genay, R. and Faruk, S. (2006). Overnight borrowing, interest rates and extreme value theory. *European Economic Review.* **50**, p. 547-563.

Jondeau, E., Poon S.H., and Rockinger, M. (2007). Financial Modeling Under Non-Gaussian Distributions. Springer.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics.* **46**, p. 323-351.

Picard P. (2003). Les frontières de l'assurabilité. *Risques.* **54**, p. 61-62.

Zajdenweber, D. (2006). Economie et Gestion de l'Assurance. *Economica*, Paris.

Correspondence Analysis in R: the ca package

Oleg Nenadić[1,*] **, Michael Greenacre**[2]
1. Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany
2. Departement d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, Barcelona, 08005 Spain
* Contact author: onenadi@uni-goettingen.de

Keywords: Multiple correspondence analysis, joint correspondence analysis, R

The `ca` package (Nenadić & Greenacre, 2007) offers functions for simple, multiple and joint correspondence analysis in the statistical software environment **R** (R Development Core Team, 2010).

Simple correspondence analysis is performed with the function `ca`, which computes the simple correspondence analysis based on the singular value decomposition. The function also allows the inclusion of supplementary points and a subset analysis. A summary method gives additional details on the perfomed correspondence analysis, such as squared correlations and contributions. The plotting functions allow for plotting the results in 2D and in 3D. They also offer a range of map scaling options.

The main function for multiple and joint correspondence analysis is `mjca`. With the option `lambda` the user can perform a multiple correspondence analysis (based on the indicator or the Burt matrix), an adjusted multiple correspondence analysis or a joint correspondence analysis (see Greenacre, 2007, for an overview). As with simple correspondence analysis the function also allows the specification of supplementary points and subset analyses. Where applicable, the structure of the multiple and joint correspondence part of the package follows the same scheme as the simple correspondence analysis part. Thus, a summary method and plotting functions in 2D and 3D are also available for multiple and joint correspondence analysis.

We present the latest revision of the package which is released at CARME 2011, where the entire part related to multiple and joint correspondence analysis has been rewritten to follow a unified approach.

References

Greenacre, M. (2007). *Correspondence Analysis in Practice, 2nd edition.* Chapman & Hall / CRC Press, London.

Nenadić, O. & Greenacre, M. (2007). Correspondence Analysis in R, with two- and three-dimensional graphics: the `ca` package. *Journal of Statistical Software*, **20(3)**.
`http://www.jstatsoft.org/v20/i03/`

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
`http://www.R-project.org/`

Principal Component Analysis of International Diffusion of Durable Goods

Javier Palacios Fenech[1,*]
1. Universitat Pompeu Fabra, Barcelona, Spain
* Contact author: xavier.palacios@upf.edu

Keywords: Diffusion, Innovations, Principal Component Analysis

Principal component analysis is used to study the interrelation of 32 consumer durable goods and 70 countries between 1977 and 2008. Countries are divided into five categories based on gross domestic product per capita at purchasing power parity. There is a natural association of countries and durable goods based on their different rates of possession. Principal component analysis reduces the dimensionality of the data, specifically, the relative position of each country regarding the consumer durable goods indicators, while keeping most of the information. The technique facilitates a visual representation of the data set in a few dimensions, which can be displayed in a single graph (biplot or principal component biplot). In the biplot, the points that represent each country in the years of the study are connected, defining the country's profiles over time. The result is a set of coordinates of each country-by-year combination and a set of coordinates of products. This visualization shows how durable goods diffuse at different rates depending on the cultural and socioeconomic characteristics of each country and it facilitates the understanding of how groups of products interact in a global framework and associate according to their different rates of possession among countries.

References

Jolliffe, I. (2002) Principal Component Analysis. Springer-Verlag, New York

Krishnan, T. V. & Thomas, S.A. (2009). International Diffusion of New Products, *The Sage Handbook of International Marketing, SAGE Publications Ltd, London.*

Peres, R., Muller, E. & Mahajan, V. (2010). Innovation Diffusion and New Product Growth Models: A Critical Review and Research Directions. *International Journal of Research in Marketing*, **27**, 91–106.

The Derivation of Individual Overall Attitude Scores from a Multiple Correspondence Analysis Solution

Marcio L.M de Souza[1], Ronaldo R. Bastos[1,*], Marcel de T. Vieira[1]

1. Departamento de Estatística – ICE/UFJF, Brasil
* Contact author: ronaldo.bastos@ufjf.edu.br

Keywords: Multivariate Analysis; Categorical Data; Multiple Correspondence Analysis; Attitude Score

Survey response data to address attitudes, satisfaction and other underlying concepts of interest to social scientists often rely on a set of Likert-type statements for which respondents choose one category among all possible categorical answers to each statement. For both exploratory and confirmatory data analyses which use a unique score to represent, for example, the overall measure of attitude of an individual, it is common to calculate such score as a summation of all values obtained from each response. However, this commonly used score is represented by integer values only and assumes equal distances between each ordered category. In addition, such summation score may be less accurate in assessing an underlying concept of interest, as two or more of these scores, although identical in value, might have come from totally different profiles. We propose a score with the intent to minimize these shortcomings.

This work proposes a simple score for each individual *i,* where raw category values (K_{ij}) for each statement response are weighted by two distinct values based on the overall solution from multiple correspondence analysis (MCA): (*a*) the inverse of the distances between each individual *i* and the category of each variable *j* to which this individual belongs ($w\mathbf{1}_{ij}$); (*b*) the inverse of the distance between the category of each variable *j* to which this individual belongs and the origin ($w\mathbf{2}_{ij}$). This score can thus be represented as $Si = \sum_J K_{ij} w1_{ij} w2_{ij} \,/\, \sum_J w1_{ij} w2_{ij}$, where the weights can be represented as $w\mathbf{1}_{ij} = [(X_i - Y_{ij}).(X_i - Y_{ij})^t]^{-1/2}$ and $w\mathbf{2}_{ij} = [(Y_{ij}).(Y_{ij})^t]^{-1/2}$. In both expressions X_i represents the score from the *n*-dimensional MCA solution for individual *i* and Y_{ij} represents the score from the *n*-dimensional MCA solution for the category of variable *j* to which individual *i* belongs.

The proposed score derivation approach was applied to attitude data from the British Household Panel Survey (see Taylor *et al.*, 2001). It was implemented in the open-source R programming language, from the MCA solution obtained through the *ca* package, *mjca* function, for a Burt matrix (Nenadic and Greenacre, 2007). In order to evaluate the stability of the results we have been undertaking simulation-based analyses with the original data and also with data generated from different population scenarios. This work presents the first results for the proposed score, which, to our view has the potential of better representing the underlying concept of interest than the mere summation of values of categorical variables over all responses.

Acknowledgement: The authors acknowledge grant CEX-APQ-00467-2008(Universal) from the Research Foundation from the state of Minas Gerais, Brasil – FAPEMIG, for the development of this work.

References

Nenadic, O. and Greenacre, M. (2007). Correspondence Analysis in R, with Two-and-Three-dimensional graphics: The ca Package. *Journal of Statistics Software*, vol. 20, issue 3. *http://www.jstatsoft.org/*

Taylor, M. F. (ed), Brice, J., Buck, N. and Prentice-Lane, E. (2001) *British Household Panel Survey - User Manual - Vol. A: Introduction, Technical Report and Appendices*. Colchester: U. of Essex.

Correspondance Analysis with Incomplete Paired Data using Bayesian Imputation

Jules J. S. de TIBEIRO [(1)(∗)] and Duncan J. MURDOCH[(2)]
(1) Université de Moncton, Moncton, N.-B., Canada
(2) The University of Western Ontario, London, ON, Canada
(∗) Contact author: jdetibeiro@stats.uwo.ca

Abstract. In this paper we consider the analysis of incomplete tables using *Correspondence Analysis* (CA). We focus on a dataset concerning congenital heart disease (Fraser and Hunter 1975), in which the data forms a square table, but only a symmetrized version of the off-diagonal entries was reported. We use *Markov chain Monte Carlo* (MCMC) on a *hierarchical Bayes model* to estimate the underlying rates, and use CA to study the relationships in the completed table.

Keywords: correspondence analysis, missing data, Markov chain Monte Carlo.

References

Benzécri, J. P. (1992). *Correspondence Analysis Handbook.* Marcel Dekker.

de Tibeiro, J. J. S. (1996). “Sur les traits associés par paires : malformations cardiaques congénitales chez des enfants ayant mêmes parents.” *Les Cahiers de l’Analyse des Données,* 21: 45-52.

Dinwoodie, I. and MacGibbon, B. (2004). “Exact Analysis of a Paired Sibling Study." *Computational Statistics,* 19: 525-534.

Dinwoodie, I. H., Matusevich, L. F., and Mosteig, E. (2004). “Transform Methods for the Hypergeometric Distribution." *Statistics and Computing,* 14: 287-297.

Fraser, F. C. and Hunter, A. D. W. (1975). “Etiologic Relations Among Categories of Congenital Heart Malformations." *The American Journal of Cardiology,* 36: 793-796.

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

Lebart, L., Morineau, A., and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices.* John Wiley & Sons.

MacGibbon, B. (1983). “A Log-linear Model of a Paired Sibling Study." In Chaubey, Y. and Dwivedi, T. D. (eds.), *Proceedings of Statistics ’81 Canada Conference,* 193-197.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Wiley. 524

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 Users Manual.* MRC Biostatistics Unit, Cambridge.

van der Heijden, P. G. M., de Falguerolles, A., and de Leeuw, J. (1989). “A Combined Approach to Contingency Table Analysis Correspondence Analysis and Log-Linear Analysis." Applied Statistics, 38: 249-292.

Gifi methods to explore EU-SILC data

Marcus Wurzer[1,*] **, Patrick Mair**[1]
1. WU Vienna
* Contact author: marcus.wurzer@wu.ac.at

Keywords: Gifi Methods, Homogeneity Analysis, EU-SILC

EU-SILC (EU Statistics on Income and Living Conditions) is an annual survey conducted in all EU member states as well as in Turkey, Switzerland, Norway and Iceland. Apart from income and living conditions, education and health are the topics that are of special interest in this survey, which is aimed at providing a basis for decision-making in social politics. The authors will present applications of various Gifi methods on EU-SILC data. Some nonstandard plots as implemented in the R package homals (de Leeuw & Mair, 2009) will be used for visualizing the results.

References

de Leeuw, J. & Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, **31(4)**, 1–21.

Analyzing multiple time series using a dynamic latent variables principal component analysis model

Simplice Dossou-Gbété

November 15, 2010

Laboratoire de Mathématiques et de leurs Applications UMR CNRS 5142. Université de Pau et des Pays de l'Adour (France)

Keywords: common trends; dynamic latent variables model; EM algorithm; Kalman filter and smoother; multivariate time series.

The statistical analysis of high-dimensional time series is an important challenge in environmental studies and hence dimension reduction is an important issue of the statistical methods involved in such studies. Therefore the detection of common patterns over time in the set of time series and relationships between these series is a central question. Most of the standard time series techniques, such as spectral analysis, wavelet analysis [3], ARIMA and Box–Jenkins models are designed for the analysis of cyclic pattern and prevision and often require stationary time series observed at equispaced time points. So they are not particularly suitable for answering above questions.

Probabilistic Principal Component Analysis (PPCA)[1] as well as Principal Component Analysis (PCA)[1] are two statistical methods designed for analyzing multivariate data. In this setting multivariate data are considered as response variables assuming latent variables (unobserved effects) could explain the variations among individual observations. These methods have proved their ability to cope with a large number of variables without running into scarce degrees of freedom problems often faced in a regression-based analysis. Similar considerations apply to multivariate time series if they are thought as response variables assuming the variations over time of the individual observations could be explained by hidden and time-varying stochastic mechanisms. These latent time-varying components could describe trends in observed time series as well as the relationships between them. This motivates the extension of Probabilistic Principal Component Analysis so as to take into account explicitly the time component that is inherent to the aims of the analysis of multivariate time series. Dynamic factor analysis is an alternative approach encountered in the literature for the analysis of the macroeconomics multivariate time series[2] as well as environmental time series [4]

This paper is devoted to a dynamic version of the probabilistic principal component analysis. Model's parameters estimation is carried out by using an implementation of the EM algorithm where the expectation step is based on Kalman filtering and smoothing. In order to show how the method works and could be helpful in investigating environmental questions, an application is carried out using a dataset that describes the behavior of a wastewater treatment plant along 527 days.

References

[1] Bishop C.M. (2006): Pattern Recognition and Machine Learning. Springer

[2] Jungbacker B., Koopman S.J. & van der Wel M. (2009): Dynamic Factor Analysis in The Presence of Missing Data.

[3] Shumway R.H. & Stoffer D.S. (2006): Time Series Analysis and Its Applications With R Examples, 2nd edition. Spîinger-Verlag

[4] Zuur A.F. et al. (2003)Estimating common trends in multivariate time series using dynamic factor analysis. Environmetrics 14, pp.665–685.

An evolutionary analysis of association patterns

Alfonso Iodice D'Enza[1,*], **Francesco Palumbo** [2]
1. Università di Cassino
2. Università degli Studi di Napoli Federico II
*iodicede@unicas.it

Keywords: Non-symmetric correspondence analysis, cluster analysis, dynamic update.

The present proposal deals with high-dimensional date sets described by several binary attributes and stratified in different subsets (or data batches) of statistical units. A typical example involving such data structures is market basket analysis (MBA) where each statistical unit is a *transaction* and the binary attributes indicate whether a product is purchased or not. Further examples are in finance, environmental and social sciences. Two main reasons may lead to a units-wise stratification: the data set is too large to be analysed in a row; the statistical units refer to different occasions in time or space. In both cases a comparison of the associations within the different data batches can be suitable. If the association analysis in high dimensional data sets can be suitably faced via factorial techniques, the comparison among different solutions obtained for each data batch, remains the main issue in the analysis. A possible solution to link the association structures of different batches is to use multiple correspondence analysis (MCA, Greenacre, 2007) of one batch and incrementally update the solution with further batches (Iodice D'Enza and Greenacre, 2010).
This paper presents an approach that, through the combination of clustering and factorial techniques, aims to study the evolution of the association structure of binary attributes over different data batches. The proposal is to introduce a latent categorical variable which is determined and updated at each incoming batch; in other words this variable is determined according to the association structure and represents the 'link' among the solutions. The latent categorical variable is endogenously determined by the procedure. The procedure consistency is assured by the fact that both the factorial technique and the determination of the latent variable satisfy the same criterion. In order to determine the latent categorical variable, a good solution consists in grouping statistical units into homogeneous groups in order to get a set of profiles that are representative of similar units.
In the literature different proposals aim to explore the relationship structure characterizing a data set through the combination of clustering procedures and factorial techniques. Procedures suitably combining clustering with factorial analysis techniques have been proposed. Vichi and Kiers (2001) propose a combination of principal component analysis (PCA) with k-means clustering method. In the framework of categorical data, another interesting approach combining clustering and multiple correspondence analysis (MCA) (Greenacre, 2007) is proposed by Hwang *et al.* (2006). Similarly, yet dealing with binary data, Palumbo and Iodice D'Enza (2010) propose a suitable dimension reduction and clustering. The present proposal is an enhancement of the latter approach to the comparative analysis of multiple batches.

References

Greenacre M. J., (2007) '*Correspondence Analysis in Practice*', second edition. *Chapman and Hall/CR.*

Hwang H., Dillon W. R. and Takane Y., (2006). 'An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents'. *Psychometrika.* 71, 161–171.

Iodice D'Enza A. and Greenacre M.J.,(2010).'Multiple correspondence analysis for the quantification and visualization of large categorical data sets'. In proc. of SIS09 *Statistical Methods for the analysis of large data-sets.* (in press).

Palumbo F. and Iodice D'Enza A.,(2010).'A two-step iterative procedure for clustering of binary sequences'. *Data Analysis And Classification.* Springer, 50–60.

Vichi M. and Kiers H., (2001). 'Factorial k-means analysis for two way data'. *Computational Statistics and Data Analysis* 37(1): 49–64.

First 50 years of Survo: from a statistical program to an interactive environment for data processing

Kimmo Vehkalahti[1,*], **Reijo Sund**[2]
1. Department of Social Research, Statistics, University of Helsinki, Finland
2. National Institute for Health and Welfare, Helsinki, Finland
* Contact author: kimmo.vehkalahti@helsinki.fi

Keywords: computing environment, editorial interface, Survo, R, Muste

Survo is an interactive computing environment for creative processing of text and numerical data. Various versions of Survo have existed during the last 50 years. The name Survo originates from the word "survey" or from the Finnish verb "survoa", meaning "to compress" (Mustonen 1992). A recently launched Muste project aims at an open source implementation of the interface and operations of Survo integrated as a part of the R project for statistical computing (`http://www.r-project.org/`).

The author of Survo is Seppo Mustonen, Professor of Statistics at University of Helsinki. Mustonen has developed and programmed the various generations of Survo, and is still responsible for further development of the current version SURVO MM, which was released 10 years ago (Mustonen 2001).

The very first Survo (in the 1960s) was a statistical program SURVO 66 running on Elliott 803 computer. In the 1970s it was followed by a Wang mini-computer version SURVO 76, which was probably the first truly interactive statistical software package in the world. In 1979, its menu-based interface was suddenly superseded by Mustonen's new innovation (which arose rather interestingly – in the context of a musical application!). The new way of working was called *editorial interface*, based on the fact that all the operations were carried out using a text editor (Mustonen 1982).

The successors of SURVO 76, namely, SURVO 84, SURVO 84C, and SURVO 98, each built on a different platform, as well as the current SURVO MM, which runs on Windows, have been based on the unique interface that Mustonen invented over 30 years ago. Through those decades, Survo has expanded in various ways and formed an integrated computing environment (Mustonen 1992, 2001).

A new Muste project (see, `http://www.survo.fi/muste/`) has been recently initiated by Reijo Sund. The aim is to create an open source implementation of the editorial interface and the operations of Survo and make them a part of the R project for statistical computing. Technically, Muste will be implemented as a fairly large R package. Since 1985, Survo has been programmed in the C language, which makes it highly compatible with the technical structure of R. In addition, Mustonen has promised to support the Muste project with all the necessary source code.

Our presentation includes examples of working with Survo and with a preliminary version of Muste. Demonstrations show, for example, how the editorial interface can be used for processing tables and matrices, making calculations, and visualising statistical data.

References

Mustonen, S. (1982). *Statistical computing based on text editing*, Proceedings of the 5th Symposium on Computational Statistics, COMPSTAT (Toulouse, France). H. Caussinus, P. Ettinger and R. Tomassone, Editors, pp. 353–358. Physica-Verlag, Wien, `http://www.survo.fi/publications/COMPSTAT_1982.pdf`.

Mustonen, S. (1992). Survo – An Integrated Environment for Statistical Computing and Related Areas, 494 pp., Survo Systems, Helsinki, Finland, `http://www.survo.fi/books/1992/Survo_Book_1992_with_comments.pdf`.

Mustonen, S. (2001). The new Windows version of Survo. Survo Systems, Helsinki, Finland, `http://www.survo.fi/mm/english.html`.

Interactive Image Mining

Annie Morin[1,*] , Nguyen-Khang Pham[1,3]

1. IRISA, Université de Rennes 1, France
3. Can Tho University, Vietnam
* amorin@irisa.frr

Keywords: Correspondance Analysis, Image Mining, Information Retrieval, Bag of words, Content based Image Retrieval, Inverted file, SIFT

We apply correspondence analysis (CA) for image mining and image retrieval . CA is very often used in Textual Data Analysis (TDA)where the contingency table crosses words and documents. In image mining, the first step is to define "visual" words in images (similar to words in texts). These words are constructed from local descriptors (SIFT, Scale Invariant Feature Transform) in images. We develop a tool CAViz which is interactive, and which helps the user interpreting the results and the graphs of CA. An application to the Caltech4 base (Sivic and al., 2005) illustrates the interest of CAViz in image mining. The method was also tested on the Stewenius and Nister datasets on which it provides better results (quality of results and execution time) than classical methods as tf*idf or Probabilistic Latent Semantic Analysis (PLSA).

Besides, to scale up and improve the retrieval quality, we propose a new retrieval schema using inverted files based on the relevant indicators of Correspondence Analysis (quality of representation and contribution to inertia). The numerical experiments show that our algorithm performs faster than the exhaustive method without losing precision. We then have extended it to build a parallel version using GPUs (graphics processing units) to gain high performance at low cost. In a large database, most time is used for filtering images . This motivates us to parallel this step. The search performance is improved by a factor of 10 in comparison to a sequential scan without loosing quality.

References

Nguyen-Khang Pham, Annie Morin, Patrick Gros, Quyet-Thang Le (2009). Accelerating image retrieval using factorial correspondence analysis on GPU. In 13th International Conference on Computer Analysis of Images and Patterns, CAIP'09, Lecture Notes in Computer science, Volume 5702, Pages 565-572, Münster, Germany.

Nguyen-Khang Pham, Annie Morin, Patrick Gros, Quyet-Thang Le (2009) . Intensive use of factorial correspondence analysis for large scale content-based image retrieval. In Advances in Knowledge Discovery and Management, AKDM'09, Springer-Verlag,.

Annie Morin, 2004, Intensive use of correspondence analysis for information retrieval in Proceedings of the 26th International Conference on Information Technology Interfaces, ITI2004, , pp. 255–258.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harsman, 1990, Indexing by latent semantic analysis, Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391–407.

T. Hofmann, 1999, Probabilistic latent semantic analysis in Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99), pp. 289–296.

D. Nister and H. Stewenius, 2006 Scalable recognition with a vocabulary tree in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 2161–2168.

J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, 2005, Discovering objects and their location in image collections in Proceedings of the International Conference on Computer Vision, pp. 370–377.

K. Mikolajczyk and C. Schmid, 2004,Scale and affine invariant interest point detectors Proceedings of IJC V, vol. 60, no. 1, pp. 63–86.

D. G. Lowe, 2004 Distinctive image features from scale-invariant keypoints in International Journal of Computer Vision, pp. 91–110.

Projective tests using Napping®, the Rorschach test revisited: are the cultural differences between Asians and Caucasians significant?

Buche Marianne[1,*], Cadoret Marine[1], Lê Sébastien[1]

1. Agrocampus Ouest, Laboratoire de mathématiques appliquées, Rennes, France
* Contact author: marianne.buche@yahoo.com

Keywords: projective test, napping®, Rorschach test, multiple factor analysis

"In psychology, a projective test is a personality test designed to let a person respond to ambiguous stimuli, presumably revealing hidden emotions and internal conflicts. (…) The best known and most frequently used projective test is the Rorschach inkblot test, in which a subject is shown a series of ten irregular but symmetrical inkblots, and asked to explain what he/she sees. The subject's responses are then analyzed in various ways, noting not only what was said, but the time taken to respond, which aspect of the drawing was focused on (…) (Wikipedia)."

The aim of this study is to revisit the Rorschach test by using the inkblots as a support and the napping® as a way to project the subject's personality on a map (tablecloth). The idea of the napping®, *aka* projective mapping, is to position a set of items on a tablecloth according to how they are perceived to be related (Pagès, 2005). Data are then analyzed using multiple factor analysis (Escofier and Pagès, J, 1988-1998) applied on groups of x-coordinates and y-coordinates for the set of inkblots, one group being associated to one subject. In addition the subject may describe the items once positioned: those descriptions are used to supplement the items' position and to enhance the interpretation of the tablecloth.

In our study, we asked two groups of 20 subjects, Asians on the one hand, Caucasians on the other hand, to perform the task previously described as we wanted to check the hypothesis of difference of perception between the two cultures. To answer that question we applied a hierarchical multiple factor analysis (Le Dien and Pagès, 2003) on the data considering first two groups of variables, the two cultures; then 40 groups of coordinates, one per subject. Such analysis allowed balancing the part of each subject within his culture as well as the part of both cultures. It allowed also comparing both cultures within one single framework.

References

Projective test, from Wikipedia, the free encyclopedia.
http://en.wikipedia.org/wiki/Projective_test.

Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis; application to the study of ten white from the Loire Valley. *Food quality and preference*. **(16)**. pp. 642-649.

Escofier, B., Pagès, J. (1988-1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*, Dunod, Paris.

Le Dien, S. & Pagès, J. (2003). Hierarchical Multiple Factor Analysis: application to the comparison of sensory profiles. *Food Quality and Preference*. 14 (5-6), 397-403.

Euclidean representations of a set of hierarchies using Multiple Factor Analysis

Marine Cadoret[1*], Sébastien Lê[1], Jérôme Pagès[1]

1. Agrocampus Ouest, Laboratoire de mathématiques appliquées, Rennes (France)
* Contact author: marine.cadoret@agrocampus-ouest.fr

Keywords: Euclidean Representations, Hierarchical Sorting Task, Hierarchy, Multiple Factor Analysis

The aim of this presentation is to propose a new approach for analyzing hierarchies issued from unsupervised classifications performed on the same (statistical) individuals. This issue has already been partially addressed by several authors for the comparison of different classifications methods (Leclerc (1985), Leclerc and Cucumel (1987) or the special issue of Journal of Classification (Vol. 3, 1986) dedicated to the comparison and consensus of classifications).

The starting point of our research framework is the hierarchical sorting task commonly used in psychology and sensory analysis (Egoroff, 2005; Qannari et al., 2010). This method consists in asking subjects to provide each their own hierarchical tree from the same given set of objects. This hierarchical tree is constructed mostly in a binary and descending way: the subjects are asked to divide the objects into two homogeneous groups and then to divide again each of the two groups until they consider the final groups homogeneous. The main feature of this method is that each subject uses his/her own criteria for making these successive divisions. In this kind of experiment were interested into getting a consensus representation of the objects from all the subjects as well as a representation of the subjects, function of the way they classified the objects.

In this talk, we propose a methodology which provides on the one hand a Euclidean representation of the objects and on the other hand a Euclidean representation of the hierarchies (i.e. a subject can be assimilated to his/her hierarchy) linked to the previous one in the manner of Multiple Factor Analysis (MFA; Escofier and Pagès, 1998). This hierarchy representation allows visualizing the different steps taken by each subject and to understand in a certain way his/her cognitive process.

The data associated with a hierarchy j can be gathered in a data table with I rows and Q_j columns (with Q_j the number of levels associated with the hierarchy j). In this case, each level of the hierarchy can be assimilated to a qualitative variable with as many modalities as there are groups for this level.

The data coming from a set of hierarchies can be gathered in a table that juxtaposes the tables associated with each hierarchy. This data table is composed of I rows and $Q = \sum_j Q_j$ columns: each row corresponds to an object and each column to a level associated with a given hierarchy; the columns are grouped by hierarchy.

This kind of table only composed of qualitative variables structured in groups (one group = one hierarchy) can be analyzed by Multiple Factor Analysis: Escofier and Pagès (1998). MFA is applied to a data table in which the same set of individuals (here the objects) are described by several sets of variables (here the hierarchies) structured in groups. MFA balances the influence of each group of variables (i.e. each hierarchy) in the analysis making maximum axial inertia of the clouds associated with the separated analysis of each hierarchy equal to 1.

This methodology will be illustrated with an example in which 24 subjects performed a hierarchical sorting task on 16 advertisements concerning an orange juice.

References

Egoroff, C. (2005). *How to measure tactile perception: a case study on automotive fabrics*, ESN Conference (Madrid, Spain), 25-26 May.

Escofier, B., Pagès, J. (1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*, Dunod, Paris.

Leclerc, B.(1985). La comparaison des hiérarchies : indices et métriques. *Mathématiques et Sciences Humaines*, **92**, 5–40.

Leclerc, B., Cucumel, G.(1987). Consensus en classification : une revue bibliographique. *Mathématiques et Sciences Humaines*, **100**, 109–128.

Qannari, E.M., Courcoux, P., Taylor, Y., Buck, D., Greenhoff, K. (2010). *Statistical issues relating to hierarchical free sorting task*, 10th Sensometrics Conference (Rotterdam, the Netherlands), 25-28 July.

Analysis of sorting data using multiple correspondence analysis and a related method

El Mostafa Qannari[1,2,3,*], Philippe Courcoux[1,2,3], Véronique Cariou[1,2,3]

1. ONIRIS, Nantes, F-44322, France.
2. INRA, Nantes, F-44316, France
3. Université Nantes, Angers, Le Mans, France

* Contact author: elmostafa.qannari@oniris-nantes.fr

Keywords: multiple correspondence analysis, sorting data.

The interest for the free sorting procedure in sensory evaluation is gaining ground as it provides a quick and reliable means to assess similarities among a set of stimuli by a panel of subjects. In this procedure, the subjects are presented with a set of stimuli and instructed to sort them in as many groups as they believe it necessary, considering that stimuli in the same group are perceived as similar. Very often, MDS techniques are used to analyse sorting data (see for instance, Faye *et al.*, 2006) but multiple correspondence analysis (MCA) was also used in this context (Cadoret, Lê & Pagès, 2009; Takane, 1981; Van der Kloot and Van Herk, 1991).

We propose a strategy of analysis of sorting data which leads to MCA. This strategy presents the advantage of explicitly showing the contribution of the subjects to the determination of the factorial axes. We also propose refinements over MCA to cope with the two well known problems relating to the impact of the variables with high numbers of categories and the impact of the "rare" categories (i.e. categories with small occurrences).

References

Cadoret, M., Lê, S. & Pagès, J. (2009). A Factorial Approach for Sorting Task data (FAST) *Food Quality and Preference*. **20**. pp. 410-417.

Faye P., Brémaud D., Teillet E., Courcoux Ph., Giboreau A. and Nicod H. (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference*, **17**, 604-614.

Takane, Y. (1981). *MDSORT*: A special purpose multidimensional scaling program for sorting data. *Journal of Marketing Research*, **18**, 480-481.

Van der Kloot W. A. and Van Herk H. (1991). Multidimensional scaling of sorting data: a comparison of three procedures. *Multivariate Behavioral Research*, 26, 4, 563-581.

Validation of ideal profile data using multivariate analysis: the ideal products' space as a link between the products and their preferences

Thierry Worch[1,2,*], Sébastien Lê[2], Jérôme Pagès[2]

1. OP&P Product research BV, Utrecht, the Netherlands
2. Agrocampus Ouest, laboratoire de mathématiques appliquées, Rennes, France

* Contact author: thierry@opp.nl

Keywords: sensory analysis, Ideal Profile Method, multivariate analysis

In sensory science, the Ideal Profile Method (IPM) refers to a particular way of using consumers to collect sensory data on a set of products (food, cosmetics, etc.) in order to improve them qualitatively. Consumers are asked 1) to rate each product according to the intensity they perceive for a list of attributes, 2) then to give an ideal score of intensity for each product tasted for the same list of attributes, 3) and finally to give a liking score. Whereas consumers are used ever since to give liking scores, it has been shown lately that they could also be used to describe products (Husson *et al.*, 2001 and Worch *et al.*, 2010); a task that was usually done by experts or trained panelists.

The aim of this presentation is to present a methodology that allows validating the consistency of the ideal sensory profiles given by consumers in the sense that if a consumer likes a product which is described as having a rather high score for an attribute, then his ideal product should rather have a high score for this attribute; as by definition, the ideal product is a product which sensory characteristics maximize the appreciation.

To do so, we first build the so called "ideal profiles" data table where rows correspond to consumers and columns to attributes they have rated: at the intersection of one row and one column, the average ideal score for a given consumer and a given attribute. For this data table, a row can also be interpreted as the ideal associated with a consumer. Thus, a principal component analysis (PCA) performed on this "ideal profiles" data table will represent two ideals products (respectively associated with two consumers) all the more close as they have been described the same way.

To that analysis, we may add the so called "sensory profiles" data table as a supplementary data table, where rows correspond to products and columns to attributes: at the intersection of one row and one column, the average score for a given product and a given attribute. The rows (products) of this data table will be projected as supplementary individuals in the space of the ideals products.

We may also add the so called "hedonic scores" data table as another supplementary data table, where rows correspond to consumers and columns to products: at the intersection of one row and one column, the hedonic score for a given consumer and a given product. The columns (products) of this data table will be projected as supplementary variables in the space of the attributes.

In this presentation, we will show in how checking the consistency of ideal profile data as defined previously consists in checking that the products represented as supplementary individuals in the space of the ideals products have the same relative positioning as the products represented as supplementary variables in the space of the attributes.

References

Husson, F., Le Dien, S., & Pagès, J. (2001). Which value can be granted to sensory profiles given by consumers ? Methodology and results. *Food Quality and Preference*, **12**, 291-296.

Worch, T., Dooley, L., Meullenet, J.F., & Punter, P.H. (2010). Comparison of PLS dummy variables and Fishbone method to determine optimal product characteristics from ideal profiles. *Food Quality and Preference*, **21**, 1077-1087.

Worch, T., Lê, S., & Punter, P.H. (2010). How reliable are the consumers? Comparison of sensory profiles from consumers and experts. *Food Quality and Preference*, **21**, 309–318.

Textual and lexical statistics

Mónica María Bécue Bertaut[1,2,*]

1. Universitat Politècnica de Catalunya. Departament de Estadística i Inv. Operativa
2. IDT.Institute of Law and Technology of the Universitat Autònoma de Barcelona

* Contact author: monica.becue@upc.edu

Keywords: textual data; textual statistics; correspondence analysis; lexicometry; constraint clustering methods

Statistics methods applied to the particular data that the texts are, in their very diverse forms, is a huge domain. In this work, we will focus on tools that belong to textual and lexical statistics.

The problems tackled are usually divided into two types: form versus content of texts. However, in fact, both aspects intertwine. A statistical approach is applied to such diverse sets of documents as classical works, political speeches, newspaper articles, collections of scientific research papers, closing speeches for the prosecutions in trials, free-answers to open-ended questions in surveys, short free-text comments in sensory data collection, etc. We can have to deal with a set of texts, or corpus, with objectives such as to detect similarities and differences, to build a partition of the texts into clusters and/or to characterize every text as compared to the others. Under other circumstances, we have to study a single text aiming at revealing its structure and evolution, that is, how the author has elaborated and organized the argumentation.

In every case, the searched information depends on the objectives and on the nature of the texts. This will drive the selection of the textual units (tool or/and full words; keeping all the words versus selecting particular words) and textual data preprocessing and coding.

Textual statistics adopt a multidimensional approach. The corpus to be analyzed is coded through a table documents×words. Correspondence analysis (Benzécri,1976; Benzécri, 1981; Lebart & Salem, 1998; Murtagh, 2005), starting from the distribution of the different words in the texts or parts of the texts, is the key method in this approach. The present possibilities of the computers increase its potentiality to visualize the information extracted from the analyzed texts. Clustering, or constrained clustering, is usually associated to correspondence analysis to enrich and complete the interpretation.

Other methods, peculiar to the textual domain and grouped under the name of lexical statistics (Muller; 1977), are also profitable to extract information from the texts. Born around the project "*Trésor de la langue française*" (Treasure of the French language) in the fifty's, these methods mainly study the richness, specificity, increase and evolution of vocabulary, that is, characteristics of the style of an author and adaptation to the circumstance of the audience and/or to the type of work.

Both groups of methods can be jointly used with profit. We will show the main results that they provide in the study of a closing speech on behalf of the prosecution in a lawsuit for murder. This speech has to prove a hypothesis, persuade and convince the audience. The strategy elaborated by the prosecutor leaves signs in the chosen words and their distribution within the text. To detect these signs allow for putting to the fore important rhetorical features. The whole of the methods help to reveal the evolution of the speech, locate the drawbacks and identify the moments of disruptions. This allows for segmenting the speech in homogeneous temporal periods that are, further, described by their characteristic words.

Other applications will be briefly mentioned to put to the fore the types of conclusions that can be drawn from statistical analyses of texts.

References

Benzécri, J.P. (1976). *L'Analyse des Données II. Correspondances*, 2nd éd., Dunod. Paris.

Benzécri (1981). *Pratique de l'analyse des données. Tome 3. Linguistique & Lexicologie.* Dunod, Paris.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring textual data*, Kluwer, Dordrecht.

Muller, Ch. (1977). *Principes et méthodes de statistique lexicale,* Paris, Hachette.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall.

Three-mode correspondence analysis: Some history and an ecological example from the sea bed

Pieter M. Kroonenberg

A short review will be given of the history of three-mode correspondence analysis starting with the rise of three-mode component models which form its core, like the singular value decomposition is the core of standard correspondence analysis (Kroonenberg, 2008). The technique will be illustrated with the data from an experiment which was conducted at the Norwegian Institute for Water Research using sediment collected from Bjrnhordenbukta, a small sheltered bay in Oslofjrd. Ninety-eight areas of homogenized sediment were subjected to one of seven levels of organic enrichment, combined with one of seven different frequencies of physical disturbance, each replicated once (Widdicombe & Austin, 2001). The effect on the biodiversity of the different levels of the factors and their interaction was examined via graphical displays resulting from three-mode correspondence analysis using the program suite 3WayPack (Kroonenberg & De Roo, 2010).

References:

Kroonenberg, P. M. (2008). Applied multiway data analysis. Hoboken, NJ: Wiley.

Kroonenberg, P. M., De Roo, Y. (2010). 3WayPack. A program suite for three-way analysis. Leiden: The Three-Mode Company.

Widdicombe, S. & Austin, M. C. (2001). The interaction between physical disturbance and organic enrichment: An important element in structuring benthic communities. *Limnology & Oceanography*, 46, 1720-1733.

The Power STATIS-ACT method

Jacques Bénasseni[1] **, Mohammed Bennani Dosse**[1,*]
1. IRMAR UMR CNRS 6625, University of Rennes 2
* Contact author: mohammed.bennani@univ-rennes2.fr

Keywords: STATIS-ACT method, power based criterion, multiset data, three-way data, RV-coefficient

The STATIS-ACT strategy is commonly used to analyse several data tables measured on the same observation units or variables. Among the successive steps involved in the method, one is devoted to finding a "compromise solution" between some inner product matrices derived from the initial tables. This compromise solution is a linear combination of the matrices optimising a given criterion. In this work, we discuss a s-power based extension of this criterion and investigate its properties. It is shown that the $s = 1$ case leads to a simplified compromise making easier the corresponding interpretations. Low rank versions of the compromise solution are also discussed and the whole results are illustrated with several real data sets and a simulation study.

References

Lavit, C., Escoufier, Y., Sabatier, R. & Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, **18**, 97-117.

Lavit, C. (1985). Application de la méthode STATIS. *Statistique et Analyse des données*, **10**, 103-116.

Vivien, M. & Sabatier, R. (2004). A generalization of STATIS-ACT strategy : DO-ACT for two multiblocks tables. *Computational Statistics & Data Analysis*, **46**, 155-171.

Representing interaction in multiway contingency tables: MIDOVA, CA and log-linear model.

Martine Cadot[1,2,*] , Alain Lelu[2,3,4]

1. Université Henri Poincaré, Nancy1
2. Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA, Nancy)
3. Université de Franche-Comté/LASELDI, Besançon
4. Institut des Sciences de la Communication du CNRS, Paris

* Martine.cadot@loria.fr

Keywords: Interaction, itemsets, loglinear model, N-way contingency table, categorical data

Correspondence Analysis (CA) is particularly suited to categorical variables, as long as 2-way contingency tables are concerned. (Mourad 1983) has pointed out that its extension to 3-way contingency tables is far from trivial, due to interaction effects between the variables. (Escofier 1983) has provided a non-symmetric solution to this problem, through the example of a 3-way *qualification*×*profession*×*gender* table, disregarding interaction in this first approach. Then (Escofier & Pagès 1988) took interaction into account, still in a non-symmetric scheme, and illustrated with the same example, and in (Abdessemed & Escofier 2000) this CA approach was contrasted with the log-linear model one.

Beside CA and log-linear model, issued from the statistics domain, other research streams originating in Artificial Intelligence have coped with the same problem: we will present here the extension to categorical variables of our results on extracting and statistically validating « itemsets » in boolean datatables, results first published in (Cadot 2006) – for a survey on itemset approaches, see (Han 2001). We coined MIDOVA (Multidimensional Interaction Differential of Variation) our method for highlighting and representing complex links between qualitative variables, which includes interaction, well-suited to socio-economic data (Haj Ali & Cadot 2010). We will compare it to the CA and log-linear model approaches, using the same 3-way example as Escofier and her colleagues. We will show that out method is effective for general N-way interactions (N may be far greater than 3), whether symmetrically or not, and results both in easy and detailed interpretability, as CA does, and in statistical significance testing, as the log-linear model does in the case of few variables.

References

Abdessemed, L. & Escofier B. (2000). Analyse de l'interaction et de la variabilité inter et intra dans un tableau de fréquence ternaire. In Moreau, J., Doudin, P.-A. & Cazes, P. (eds). *L'analyse des correspondances et les techniques connexes*, 146-164. Springer-Verlag, Berlin.

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Ph.D. thesis, Université de Franche-Comté, France.

Escofier, B. (1983). *Généralisation de l'analyse des correspondances à la comparaison de tableaux de fréquences*. Rapport de Recherche Inria, Rennes, N°207.

Escofier, B. & Pagès, J. (1988). *Analyses factorielles simples et multiples.* Dunod Paris.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco.

Haj Ali, D. & Cadot, M. (2010). *Estimation de l'impact de la décision du mariage sur la pauvreté des ménages tunisiens*, MASHS 2010 (Lille, France), 10–11 juin, pp. 45–56

Mourad, G. (1983). Flux de pétrole et flux de marchandises entre l'OPEP et l'OCDE de 1970 à 1979. In Benzécri J.-P. & collaborateurs (eds). *Pratique de l'analyse des données*, tome 5, économie, pp. 233–280.

Canonical correspondence analysis for uncovering temporal features in chronological textual data

Belchin Adriyanov Kostov[1,*], Mónica Bécue-Bertaut[2,3], Annie Morin[4]

1. Gesclinic_CAP Les Corts, Barcelona (Spain)
2. Departament Estadística I Investigació Operativa, Universitat Politècnica de Catalunya, Barcelona (Spain)
3. Institut Dret i Tecnologia, Universitat Autonòma de Barcelona, Sardanyola (Spain)
4. IRISA, Université de Rennes 1, Rennes Cedex 35042 (France)

* Contact author: belchin86@hotmail.com

Keywords: Textual data, Correspondence analysis, Canonical correspondence analysis, Chronological clustering.

In this work, we focus on chronological textual data, such as chronological series of newspaper articles, bibliographical data collected for performing technology watch on a given topic, etc. The objectives are to identify words (or association of words) that:

- are gradually renewed –mark of the time flux, usually observed in any kind of chronological corpus (continuous temporal evolution or chronology);
- or present cyclical occurrences (seasonality);
- or are recurrent but without any temporal pattern (irregular recurrent event)
- or concern a unique moment (specific event that upsets the discourse flow).

We also want to identify the different kind of events that have been encrypted in the discourse.

Correspondence analysis (CA) is a reference method to deal with textual data (Benzécri, 1981; Lebart et all., 1998). However, due to the richness and complexity of the textual data, the results can be difficult to read and interpret. Canonical correspondence analysis (CCA; ter Braak, 1986) provides more interpretable graphics by introducing explicative variables in the analysis.

Presented and classically used in ecology field, this statistical method analyses a multiple table gathering both a sites×species frequency table and a sites×environmental variables table. CCA looks for the dispersion directions of the sites and species, in a CA-like way, but constrained to be linear combinations of the environmental variables (quantitative or categorical) that play an explicative role.

In the kind of chronological textual data of our interest, we propose to apply CCA to both documents×words and documents×explicative_variables tables, being the explicative variables the chronology (eventually discretized but keeping short intervals of time) and the season (categorical variable with as many categories as convenient). Further, a chronological clustering (Legendre & Legendre, 1998) allows for identifying homogeneous time periods and discontinuities.

We illustrate the methodology by means of collections of articles extracted from the French newspaper *Le Monde*. An analysis has been performed on the articles corresponding to three sections (*international* from 1987 to 2003, *sport* from 1995 to 2003, *sciences* from 1995 to 2003), being the year (chronology) and the month (seasonality) the explicative variables. The articles of a same month are gathered into one document. Chronology is dominant in *International* section while cyclical events characterize *sport* (in particular, the *Tour de France* every July). The variations in S*ciences* section look to be explained neither by chronology nor seasonality.

References

Benzécri, J.P., 1981. *Pratique de l'analyse des données*, Vol. 3, Linguistique & Lexicologie. Dunod, Paris.

Lebart, L., Salem, A., Berry, L. (1998). *Exploring textual data*, Kluwer, Dordrecht.

Legendre P., Legendre L. (1998). Numerical Ecology. Elsevier Science B.V., Amsterdam.

ter Braak, C.J.F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167-1179.

Excellent news for German universities? A Multiple Correspondence Analysis of media reporting on the Excellence Initiative

Stefan Hornbostel[1,2], Christoph Marty[1,*]

1. Institute for Scientific Research and Quality Assurance (iFQ), Bonn

2. Department of Social Sciences, Humboldt University Berlin

* contact author: marty@forschungsinfo.de

Keywords: Media Sociology, Content Analysis, Excellence Initiative, Rhetoric of Excellence, Bourdieu

The "Excellence Initiative" (ExIn) – a research funding program of both the German Federal Government and the 16 Federal States – aims at increasing the "international visibility" of German universities. Thus, it breaks with the postulate of all universities' equality and instead gives an impulse for more differentiation in the university system. This readjustment of German science policy is mirrored by the metaphor "lighthouse of science". The promise for this new type of academic reputation has initiated a severe competition among German universities, culminating in unofficial titles like "Elite-University" that signalize explicitly distinction from others. In our opinion, the ExIn thus produces symbolic capital according to the sociology of Bourdieu. The appreciation of this symbolic capital results from the public attention on the ExIn which is generated by intense media coverage.

The ExIn gains its legitimacy from the decision-making of internationally appointed panels of experts. But in contrast to public expectations aroused by a diffuse Rhetoric of Excellence, the peer review based approvals of ExIn are controversial, e.g. because of a lack of reliable performance indicators. The reporting on ExIn in the press mirrors this collision of scientific judgments' fragility and public reasoning over quality in science. In order to sharpen the diffuse term "Rhetoric of Excellence" and learn more about its implied expectations on science, we performed a content analysis of media coverage on ExIn. We use Multiple Correspondence Analysis (MCA), which has proven to be an adequate method in media sociology (Schäfer 2008: 391), for revelation of the structure of the discourse by describing thematic and evaluative differences in the newspapers' reporting.

A computer-aided content analysis in the press data base GENIOS revealed five peaks in media coverage on ExIn: its resolution by politics (summer 2005) and the announcements of the decisions in the two preliminary and final rounds (January 2006 & 2007, October 2006 & 2007). Based on a quantitative category system, we analyzed a total of 580 news articles, which were published in defined time-spans before or after these events in one of Germanys top-5 high circulation national papers (Süddeutsche Zeitung, Frankfurter Allgemeine Zeitung, Die Welt, Franfurter Rundschau, die tageszeitung), the weekly paper Die Zeit or the local paper Tagesspiegel For each article (unit of analysis) it was registered, whether "hot topics" of public debate was discussed or not. Additionally, the evaluative tone was determined (one variable with "positive", "negative" or "neutral"). A test for intracoder-reliability was successfully performed by retesting 10 % of the articles using the same variables.

MCA bases on an adjusted Burt matrix and was performed for the whole sample and – in order to identify changes in media coverage over time - separately for the five defined time spans. The resulting maps visualize differences between the newspapers. Firstly, the media differ from each other in their evaluation of the ExIn. Secondly, the setting of topics varies. Altogether, these impressions provided by MCA give a new perspective on the discourse on ExIn and the nature of the accompanying Rhetoric of Excellence.

References:

Schäfer, Mike S. (2008): Diskurskoalition in den Massenmedien. Ein Beitrag zur theoretischen und methodischen Verbindung von Diskursanalyse und Öffentlichkeitssoziologie. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 60(2), S. 367-397

These are first results from the iFQ-project „Science & Media: Fragile and conflicting scientific evidence in the decision-making process of the Excellence Initiative and its Media Coverage", which is funded by the German Research Foundation (DFG) in the Priority Program SPP 1409 "Science and the General Public: Understanding Fragile and Conflicting Scientific Evidence."

Confidence ellipses when analyzing simultaneously several contingency tables resulting from free-text descriptions

Cadoret Marine[1,*], Buche Marianne[1], Lê Sébastien[1]

1. Agrocampus Ouest, laboratoire de mathématiques appliquées, Rennes, France
* Contact author: marine.cadoret@agrocampus-ouest.fr

Keywords: confidence ellipse, correspondence analysis, multiple factor analysis, free-text description

Textual data are often analysed using correspondence analysis (CA) on a contingency table where, for instance, rows correspond to the texts of the corpus to be analysed and columns to the words used in the corpus: at the intersection of one row and one column there is the number of occurrences of the word (associated with the column) in the text (associated with the row) (Lebart, Salem & Berry, 1997).

In our particular application, a given set of items is described by two groups of subjects using free-text description. For each group of subjects we may build a contingency table where rows correspond to the items to be studied and columns to the words used to describe the items; we may then obtain a representation of the items per group using CA. The comparison of those two representations may be obtained by the simultaneous analysis of both contingency tables using the so called intra-sets multiple factor analysis (Bécue and Pagès, 1999; Escofier, Pagès, 1988-1998). This method provides a global representation of the rows as well as a partial representation of the rows from the point of view of each contingency table, within a single framework.

For each of those different representations, global and partial, we may wonder what might have been the positions of the items if the description had been generated by some other subjects. To answer that question, we propose a methodology that allows building confidence ellipses around the items that would represent the variability of the positions the items might have taken for other subjects (Lê, Husson & Pagès, 2004).

To build such ellipses, the idea is to resample the subjects with replacement and to build from those particular subjects' description a contingency table to be projected as supplementary elements on the axis issued from the analysis of the original groups of subjects. We then obtain a new representation of the set of items from a virtual group of subjects. Ellipses are finally obtained after having resampled a great number of times.

This methodology will be illustrated using as items the Rorschach's inkblots, two groups of subjects, a first one that has analyzed the cards following the official order of the Rorschach's test, a second one that has analyzed the cards following a random order.

References

Bécue, M., Pagès, J. (1999). Intra-Sets Multiple Factor Analysis. Application to textual data. *Proc. of the 9th International Symposium on Applied Stochastic Models and Data Analysis*, J. Jansen *et al.* (eds), Universidade de Lisboa Editor, 51-60.

Escofier, B., Pagès, J. (1988-1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*, Dunod, Paris.

Lebart, L., Salem, A., Berry, L. (1997). *Exploring textual data*, Kluwer.

Lê, S., Husson, F. & Pagès, J. (2004). Confidence ellipses in HMFA applied to sensory profiles of chocolates. The 7th Sensometrics meeting, Davis (USA).

Combinatorial Inference in Geometric Data Analysis : typicality test.

Brigitte Le Roux[1,2,*], **Solène Bienaise** [1,**]

1. Université Paris Descartes and CEREMADE, Université Paris Dauphine
2. CEVIPOF, Sciences Po Paris

* Brigitte.LeRoux@mi.parisdescartes.fr
** bienaise@ceremade.dauphine.fr

Keywords: GDA, Permutation Test, Bootstrap, Typicality test

In this paper, we present a statistical inference method for Geometric Data Analysis (GDA), that is not based on random modeling, but on permutation procedures recast in a combinatorial framework. The method is applicable to any Individuals × Variables table, with structuring factors on individuals, and either numerical (principal component analysis) or categorized (multiple correspondence analysis) variables. We outline permutation testing on the target paradigm, bringing an answer to the typicality problem.

References

Cox D.R., Hinkley D.V. (1974). *Theoretical statistics*. London : Chapman and Hall.

Cramér H. (1946). *Mathematical Methods of Statistics*, Princeton : Princeton University Press.

Edgington E. (1987). *Randomization tests*, New-York : Dekker.

Le Roux B., Rouanet H. (2004). *Geometric Data Analysis : From Correspondence Analysis to Structured Data Analysis*, Dordrecht : Kluwer.

Lindley D. (1965) *Introduction to Probability and Statistics from a Bayesian viewpoint* (Part 2), Cambridge : Cambridge University Press.

Rouanet H., Lecoutre B. (1983). Specific inference in ANOVA : From significance tests to Bayesian procedures, *British Journal of Statistical and Mathematical Psychology*, 36, 252-268.

Le Roux B., Rouanet H. (2010) *Multiple Correspondence Analysis*, series : QASS vol 163, CA : Thousand Oaks, Sage Publications.

Screening the Data for Detecting Methodological induced Variation

Jörg Blasius

Institute of Political Science and Sociology, University of Bonn

Keywords: Data screening, Categorial principal component analysis, Subset multiple correspondence analysis, social sciences

Responses to a set of items in survey data are associated with different kinds of response styles, such as acquiescence response style, extreme response style, and midpoint responding. Further, there are misunderstandings of questions, duplicates of large parts of questionnaires, arbitrary responses, fatigue and other effects, which also reduce the quality of data. In general, when analyzing a battery of items, responses are related to the substantive concept, in which social scientist are mainly interested in, and to methodological effects. Applying subset multiple correspondence analysis (Greenacre and Pardo, 2006) allows to asses the structure of subsets of the items, for example, the non-substantive or the extreme response categories. Applying categorical principal component analysis (CatPCA) to an item battery of survey data allows us to assess what part of the responses is due to substantive relationships and what part is attributable to methodological artifacts. In a first paper, Blasius and Thiessen (2009) demonstrated that the share of tied data in CatPCA can be used as a rough indicator for assessing the quality of data. This idea has been further developed so that we are now able to provide with a coefficient to describe the quality of responses in a given item set. Using different examples, we will show which part of variation can be explained by the substantive concept and which part is due to methodological induced variation.

References

Blasius, Jörg and Victor Thiessen (2009). Facts and Artifacts in Cross-National Research: The Case of Political Efficacy and Trust. In: Max Haller, Roger Jowell and Tom W. Smith (eds.), *The International Social Survey Programme, 1985-2009. Charting the Globe*. London: Routledge, pp. 147-169.

Greenacre, Michael and Rafael Pardo (2006). Multiple Correspondence Analysis of Subsets of Response Categories. In: Michael Greenacre and Jörg Blasius (eds.), *Multiple Correspondence Analysis and Related Methods*. Boca Raton: Chapman & Hall, pp. 197-217.

.

Complex Sampling Designs and Multiple Correspondence Analysis

Augusto C. Souza[1,2,*], Ronaldo R. Bastos [2], Marcel de T. Vieira [2]

1. CEDEPLAR / UFMG, Belo Horizonte – MG, Brasil
2. Departamento de Estatística, ICE/UFJF, Juiz de Fora – MG, Brasil

* Contact author: augusto.ralph@gmail.com; augusto@cedeplar.ufmg.br

Keywords: Multiple Correspondence Analysis, Complex Sampling Design, Correspondence Analysis, Inference.

Issues arising from complex sampling designs for all methods of data analysis related to Correspondence Analysis (CA) are becoming increasingly important, as simple random sampling is rarely used in the process of data collection for the social sciences, and therefore much remains to be formally sorted (see, e.g. Nyfjäll, 2002).

Complex sampling designs of a finite population may consider different probabilities of object selection, stratification and clustering, not to mention other adjustments that are often made. Data thus collected must be analysed accordingly, lest unwanted non-sampling errors are unknowingly introduced in the analysis (Lehtonen & Pahkinen, 1996).

Intuitively, we can accept that if the observed raw contingency table cell frequencies are not unbiased point estimates for the underlying population cell frequencies, as is the case with most data arising from complex sampling, CA may generate maps which do not reflect the true population relationships. However, if the cell frequencies in which sampling weights have been used to "expand" the observed cell frequencies, the structural relation between lines and columns, revealed by CA applied to sample data, correctly reflects the population structure, as shown, for example, by Nyfjäll (2002), for the case of simple correspondence analysis. As CA methods are essentially descriptive in nature, the factor projections of points (profiles) in CA maps are best obtained from such "expanded" contingency tables whenever complex sampling is used to obtain the data.

As multiple correspondence analysis (MCA) presents the same algebraic features of CA, the best point estimates for the location of profiles, under complex sampling design, are to be obtained from data that have been weighted accordingly. The question that motivates this work is exactly how to incorporate such sampling weights in MCA.

MCA uses a rectangular indicator matrix (Z), where objects are represented in the lines and variable categories in the columns, with all responses coded as dummy variables. We propose to substitute the corresponding weight of each response for the original "1" values so as to generate unbiased point estimates for the profiles. In order to validate our proposal, we used the Burt matrix (B), a transformation of Z ($B=Z^TZ$), which generates a square symmetric matrix, made up of all two-way cross-tabulations of the original data set (Greenacre & Blasius, 2006).

As the Burt matrix is simply an alternative data structure for MCA, the solutions obtained by either method are necessarily identical. Our argument, therefore, is that the input of one matrix cannot differ from the input of the other, considering that B can be obtained from Z. So, since B is made up by contingency tables which have been "expanded" by the sampling weights – what one expects to lead to unbiased point estimates of profile projections – Z associated to this particular B should also present the same results and the same properties regarding the location of profiles on the solution map, capturing the effects of the complex sampling used. Although it is not possible to algebraically derive Z from B, the two matrices are equivalent in terms of the MCA geometric solution; so we assumed that the aforementioned argument is valid. Therefore, we simply calculated and compared the results obtained from both alternative ways.

The comparison we made shows that the "expanded" matrix Z correctly generates the expected B. Moreover, the algebraic results are identical for both ways. As the "expanded" B represents the best estimates for the population totals, Z adjusted by sampling weights is also the best estimate for the population Z. As a result, we propose that in order to incorporate complex sample designs to MCA solutions one must multiply each line of the indicator matrix Z by its corresponding object sample weight.

References

Greenacre, M., BLASIUS, J. (2006). *Multiple Correspondence Analysis and Related Methods*, Boca Raton: Chapman & Hall/CRC.

Lehtonen, R., Pahkinen, E. J. (1996). *Practical Methods for design and analysis of complex surveys*, Revised edition, John Wiley & Sons.

Nyfjäll, M. (2002). *Aspects on Correspondence Analysis Plots under Complex Survey Sampling Design*, Research report 2002:2 Departament of Information Science, Division of Statistics. Uppsala University.

Correspondence Analysis of Surveys with Conditioned and Multiple Response Questions

Amaya Zárraga[1,*] **, Beatriz Goitisolo**[1]

1. Departamento de Economía Aplicada III. UPV/EHU. Bilbao. Spain * Contact author: amaya.zarraga@ehu.es

Keywords: Correspondence Analysis, Multiple Correspondence Analysis, Complete Disjunctive Tables, Incomplete Disjunctive Tables

Correspondence Analysis (CA) of surveys studies the relationship between several categorical variables defined with respect to a certain population. However, one of the main sources of information are those surveys in which it is usual to find multiple response questions and/or conditioned questions that do not need to be answered by the whole population. In these cases, the data codified as 0 (category of no chosen response) and 1 (category of chosen response) can be expressed by means of an incomplete disjunctive table (IDT). The direct application of standard CA to this type of table could lead to inappropriate results. In order to apply classical CA the data can be codified in a complete disjunctive table (CDT). But this requires the inclusion of "fictitious" categories that have the same importance in the analysis than the responses of individuals and they even can create the first factors. We therefore propose a methodology for the analysis of surveys with conditioned and multiple response questions.

References

Escofier, B. (1987). Traitement des questionnaires avec non-réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. *Publications de l'Institut de Statistique de l'Université de Paris*, **XXXII(fasc 3)**, 33–70.

Escofier, B., & Pagès, J. (1998). *Analyses Factorielles Simples et Multiples. Objectifs, Méthodes et Interprétation.* 3e édition, Dunod, Paris.

Lebart, L., Piron, M. & Morineau, A. (2006). *Statistique exploratoire multidimensionnelle : visualisations et infrences en fouille de donnes.* 4e édition, Dunod, Paris

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

Zárraga, A., & Goitisolo, B. (1999). Independence between questions in the factor analysis of incomplete disjunctive tables with conditioned questions. *Questiió*, **23(3)**, 465–488.

Zárraga, A., & Goitisolo, B. (2008). Análisis de encuestas con preguntas condicionadas. *Metodología de Encuestas*, **10**, 39–58.

Constructing a Socio-Economic Status Index for a Non-Homogenous Society with Distinct Sets of Variables in Multiple Correspondence Analysis

Sugnet Lubbe[1*], Sheetal Silal[1], Niël J le Roux[2]

1. Department of Statistical Sciences, University of Cape Town, South Africa
2. Department of Statistics and Actuarial Science, Stellenbosch University, South Africa

* Contact author: Sugnet.Lubbe@uct.ac.za

Keywords: Multiple Correspondence Analysis, Socio-Economic Status

Multiple correspondence analysis (MCA) is frequently used for the visualisation of social survey data. In a set of variables associated with socio-economic status (SES) it is expected that there is some positive correlation between the variables and that the first MCA component can act as an index or ordering of SES. In this paper we will concentrate on constructing a SES index based on several such sets of variables. In particular, the data set obtained from the Researching Equity in Access to Health Care project in South Africa deals with the reality of the South African situation of merging different perspectives on SES. The naïve combination of variables without taking into account differences in the developed and developing components of a mixed society can have the opposite effect to the intention of supplementing each other into a combined measure. The difficulties in merging diverse sets of variables to construct a SES index in a mixed society will be explored and discussed.

Deliberate Self Harm among Irish Adolescents

Andrew Grannell[1] , Dr. Tony Fitzgerald[1,2,*], Dr. Paul Corcoran[3]

1. School of Mathematical Sciences, University College Cork, Ireland
2. Department of Epidemiology and Public Health, University College Cork, Ireland
3. National Suicide Research Foundation

* Contact author: t.fitzgerald@ucc.ie

Keywords: Multiple Correspondence Analysis, Biplot, Deliberate self-Harm, Irish Adolescents

Deliberate self harm (DSH) is widely recognised as a major public health issue among adolescents in Ireland (Morey et al, 2008; Keely, H, 2004; McMahon et al, 2010). Currently, adolescents in Ireland are considered to be at the highest risk, with young women having the highest number of cases admitted to A&E departments for DSH. Due to the lack of information and data on the topic, an international study was conducted across seven countries (Australia, Belgium, England, Hungary, Ireland and Holland) (Madge et al, 2008).

The aims of our study are to (1) analyse specific factors associated with DSH using Multiple Correspondence Analysis (MCA) to investigate what levels of these factors are associated with various levels of DSH and (2) investigate the differences, if any, between males and females with respect to the specific factors and DSH.

The instrument incorporated in the survey conducted by Madge et al (2008) was an anonymous, self-completed questionnaire which students had 30 minutes to answer as to facilitate its completion within one class period at school. Among the various aspects covered in this survey, three validated psychological scales were used to gain insight into depression, anxiety, self-esteem and impulsivity amongst adolescents in Ireland (McMahon et al, 2010). We incorporated the geometric approach to MCA, which deals with the visualisation of data, in our study.

When analysing various groupings of factors associated with DSH, it was evident that certain levels of factors, be it physical abuse or psychological characteristics, have different associations with the different levels of DSH. It was also shown that slight differences exist between males and females when looking only at the specified factors. When these factors are examined more closely, and each level of these factors is taken into account, the differences become more apparent. It was also interesting to observe that, while using a completely different method of analysis, the overall findings are similar to comparable studies conducted in Ireland and abroad. Multiple correspondence analysis allowed us to examine associations between variables without the use of a specified model.

References

Keeley, H. (2004). *Deliberate Self Harm in Teenagers*. 3TS Conference on Suicide in Modern Ireland, New Dimensions, New Responses (Dublin, Ireland), November 12th -14th.

Madge, N., Hewitt, A., Hawton K., Jan De Wilde, E., Corcoran, E., Fekete, S., Van Heeringen, K., De Leo, D., Ystgaard, M., (2008). Deliberate Self Harm within an International Community Sample of Young People: Comparative Findings from the Child and Adolescent Self-Harm in Europe (CASE) Study. *Journal of Child Psychology and Psychiatry*, **49(6)**, 667-677.

McMahon, E., Reulbach, U., Corcoran, P., Keeley, H., Perry, I., Arensman, E., (2010). Factors Associated with Deliberate Self Harm among Irish Adolescents. *Psychological Medicine*, **40(11)**, 1811-1819.

Morey, C., Corcoran, P., Arensman, E., Perry, I., (2008). The Prevalence of Self Reported Deliberate Self Harm in Irish Adolescents. *BMC Public Health*, **8(79)**, 1-7.

Biplots: Taking Stock

John Gower
Department of mathematics
Walton Hall
The Open University
Milton Keynes, Mk7 6AA, U.K.

It may be argued that biplots have been with us for at least 350 years but it is only since Ruben Gabriel's 1971 paper, and the pervasive availability of the VDU screen, that they have entered the modern era. In the past year two books have been published on biplots.

The basic idea is very simple: simultaneously to display graphically two kinds of entity, in such a way that gives a good visual impression of data presented in a matrix X. However, there are many questions that need investigation and whose current status will be discussed. These include:

What kinds of data are permitted in X and how do these affect their representation?
Does X require some initial transformation?
How are the two kinds of entity presented: as lines and/or points?
Usually approximation is involved: how is the quality of the display assessed and presented?
How are the diagrams to be interpreted: using inner-products, distances, angles, areas, projections?
What freedom is there in the relative scaling of the two sets of entities?
What freedom is there in the placement and orientation of lines and points?
What special considerations pertain to large data-sets and how may they be addressed?

Different answers to these questions often convey precisely the same information but presented in different graphical forms. There is no "correct" choice, making it essential for proper interpretation that users are fully aware of what choices have been made.

Advances in Visualizing Categorical Data

Michael Friendly[1,*], **Heather Turner**[2], **David Firth**[2], **Achim Zeileis**[3]
1. York University, Canada
2. University of Warwick, UK
3. Universität Innsbruck, Austria
* Contact author: friendly@yorku.ca, http://datavis.ca

Keywords: mosaic displays, generalized linear models, 3D mosaics, RC models, biplot

At CARME 1995 in Cologne, I described my work on graphical methods for visualizing categorical data, with emphasis on mosaic displays and related methods. In this talk I survey some of the advances on this topic by myself and others that have occurred over the intervening 15 years.

I illustrate these new methods and extensions using a variety of R packages. In particular: mosaic-like displays have been generalized to a wide class of graphical methods subsumed under the strucplot framework in the vcd package (Meyer *et al.*, 2009, 2006); traditional loglinear models and their generalized linear model equivalents have been extended in the gnm package (Turner and Firth, 2009) to generalized nonlinear models, providing biplot and SVD views in some cases; the vcdExtra package provides extended examples of some of these, as well as a new 3D implementation of mosaic displays.

References

Meyer, D., Zeileis, A., and Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, 17(3), 1–48. URL http://www.jstatsoft.org/v17/i03/.

Meyer, D., Zeileis, A., and Hornik, K. (2009). *vcd: Visualizing Categorical Data*. R package version 1.2-7.

Turner, H. and Firth, D. (2009). *Generalized nonlinear models in R: An overview of the gnm package*. URL http://CRAN.R-project.org/package=gnm. R package version 0.10-0.

The Mixed Effect Trend Vector Model

Mark de Rooij
Methodology and Statistics Unit, Psychological Institute, Leiden University
* Contact author: rooijm@fsw.leidenuniv.nl

Keywords: Biplots; Categorical data; Longitudinal data; Gauss-Hermite quadrature; Multilevel model.

Maximum likelihood estimation of mixed effect baseline category logit models for multinomial longitudinal data can be prohibitive due to the integral dimension of the random effects distribution. We propose to use multidimensional scaling methodology to reduce the dimensionality of the problem. As a by product readily interpretable graphical displays representing change are obtained. After formulating our generic model, we present special cases for ordinal and nominal data. Relationships to standard statistical models for multinomial data will be presented. Several empirical examples will be given to show the merits of the proposed modeling framework.

New pictures for correlation structure

Jan Graffelman[1,*]
1. Department of Statistics and Operations Research, Universitat Politècnica de Catalunya
* Contact author: jan.graffelman@upc.edu

Keywords: principal component analysis, principal factor analysis, interpretation function.

There are many ways to make a graphical representation of the correlations between a set of variables. A standard way to visualize the correlation between a pair of variables is the scatter plot, where the correlation is related to the degree of scatter around a straight line. For sets of more than two variables, biplots (Gabriel, 1971; Gower & Hand, 1996) are often used to represent correlations. Some other alternatives exist: the pictorial representations called *corrgrams* developed by Friendly (2002), and the *correlation diagrams* described by Trosset (2005). The correlation diagram represents each variable by a unit norm vector in a circle, choosing the angles such that their cosines approximate the sample correlations as well as possible. In biplots obtained by principal component analysis (PCA) there are two ways to read off a correlation: by evaluating the cosine of an angle, or by evaluating the scalar product between two vectors. In the full space of a PCA solution the cosine equals the sample correlation exactly, but it is not clear to what extent two-dimensional solutions are optimal in approximating correlations. In fact, the approximation of the correlation by the scalar product is usually better. Scalar products are more flexible, because both angle and vector lengths can be adjusted to fit to the correlations. However, the formula for the sample correlation coefficient bears a striking relationship to the trigonometric formula for the cosine of the angle between two vectors:

$$r(x,y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}}, \qquad \cos\alpha = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}. \tag{1}$$

It is therefore no surprise that cosines between angles are widely used to infer correlations. In fact, the principle pervades multivariate analysis: it is used biplots, in factor loading diagrams, canonical correlations correspond to the cosine of the angle between two linear subspaces, and so on. Many mathematicians and statisticians regard the relationship between cosine and correlation as natural and nice, and do not question it. In practice, it is pretty difficult to estimate a correlation coefficient with reasonable precision by just looking at a biplot. Moreover, there is no strict need to represent correlations by cosines. We might as well choose to represent a correlation by the *sine* of the angle if we would wish to. More generally, we can introduce a specific *interpretation function* to describe the relation between angle and correlation.

It is, as will be shown, fairly straightforward to construct a plot that represents two variables and shows their correlation in the way specified by the interpretation function. The real challenge is to do this for more than two variables. In this contribution, we will discuss several approaches to obtain multivariate plots that show correlations according to some sensible interpretation function.

References

Friendly, M. (2002). Corrgrams: exploratory displays for correlation matrices. *The American Statistician*, **56(4)**, 316–324.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58(3)**, 453–467.

Gower, J.C. & Hand, D.J. (1996). *Biplots*. Chapman & Hall, London.

Trosset, M.W. (2005). Visualizing correlation. *Journal of Computational and Graphical Statistics*, **14(1)**, 1–19.

Logistic Biplots for Binary, Nominal and Ordinal Data

José Luis Vicente-Villardón

Departamento de Estadística. Universidad de Salamanca. Spain
villardon@usal.es

Keywords: Logistic Biplot, Binary, Nominal and Ordinal Data.

Classical Biplot methods allow for the simultaneous representation of individuals and continuous variables in a given data matrix. When variables are binary, nominal or ordinal, a classical linear biplot representation is not suitable. We propose a linear biplot representation based on logistic response models. The coordinates of individuals and variables are computed to have logistic responses along the biplot dimensions. The method is related to logistic regression in the same way that Classical Biplot Analysis (CBA) is related to linear regression, thus we refer to the method as Logistic Biplot (LB). In the same way as Linear Biplots are related to Principal Components Analysis, Logistic Biplots are related to Latent Trait Analysis or Item Response Theory. The geometry of those kinds of biplots is studied: For nominal data, the linear biplot results in a partition of the representation the divides the space onto a prediction region for each category; for ordinal data, we obtain a prediction direction with points separating each category.

The usefulness of the proposal is illustrated using data on SNPs (Single Nucleotide Polymorphisms) from the HAPMAP project.

References

BAKER, F.B. (1992): *Item Response Theory. Parameter Estimation Techniques*. Marcel Dekker. New York.

GABRIEL, K. R. (1998). Generalised bilinear regresión. *Biometrika*, 85: 689 – 700.

GOWER, J. C. & HAND, D. (1986): *Biplots*. Chapman & Hall. London.

DEMEY, J., VICENTE-VILLARDON, J. L., GALINDO, M.P. & ZAMBRANO, A. (2008) Identifying Molecular Markers Associated With Classification Of Genotypes Using External Logistic Biplots**.** *Bioinformatics*, 24(24):2832-2838.

VERBOON, P. & HEISER, W. J. (1994). Resistant Lower Rank Approximation of Matrices by Iterative Majorization. *Computational Statistics & Data Analysis.* 18: 457-467.

VICENTE-VILLARDON, J. L., GALINDO M. P. & BLAZQUEZ, A. (2006). Logistic Biplots. In "*Múltiple Correspondence Análisis And Related Methods*". Grenacre, M & Blasius, J. Eds. Chapman and Hall. Boca Ratón.

Predictive nonlinear biplots: maps and trajectories

Karen Vines
Department of Mathematics and Statistics, The Open University, Milton Keynes, UK
s.k.vines@open.ac.uk

Keywords: Nonlinear biplots, normal projection, prediction, prediction regions, predictive trajectories

When the difference between samples is measured using a Euclidean-embeddable dissimilarity function, observations and the associated variables can be displayed on a nonlinear biplot (Gower and Harding, 1988). Furthermore, a nonlinear biplot is predictive if information on variables is added in such a way that it allows the values of the variables to be estimated for points in the biplot.

I will introduce a predictive nonlinear biplot map, an r dimensional plot which displays the predicted value of a variable for every point in the plot. Using such maps I will show that when the dissimilarity function with respect to a new point is not smooth everywhere, the set of predicted values can appear discrete even though the data are assumed to be continuous. That is, on an r dimensional biplot the region of points that predict a given value might also be r dimensional. Prediction trajectories that approximate 2 dimensional predictive regions are also introduced. These prediction trajectories allow information about two or more variables to be displayed on such 2 dimensional biplots.

Reference

Gower, J.C. and Harding, S.A. (1988). Nonlinear biplots. *Biometrika*, **75**,445-455.

Multiple factor analysis to two-way contingency table to compare residential and geographical trajectories

Elisabeth Morand[1*], Bénédicte Garnier [1], Catherine Bonvalet [1]

1. Institut National d' Etudes Démographiques (Ined)-Paris
* Contact author: elisabeth.morand@ined.fr

Keywords: qualitative harmonic analysis, multiple factor analysis,

Survey "Peuplement et Dépeuplement de Paris" (1986) ask a cohort Paris region inhabitants for all the homes they have leaving in the Paris region. Respondents are interviewed retrospectively about the characteristics of different units they held (including tenure and location).

The aim of this presentation is to compare residential trajectories (changes in tenure status during life) and geographical trajectories using the method of Qualitative Harmonic Analysis and method of multiple data set comparison. The Qualtitative Harmonic Analysis to study each course separately: first, the residential trajectory (tenure status) and other geographical trajectory (location: Paris, inner suburb, outer suburbs). Then the implementation of a Multiple Factor Analysis, to use both sets of variables of different types (categorical, frequency) will be used to compare both trajectories and to observe links between residential and geographical trajectories.

The method of comparison data table used was established by Becue and Pages (2008). The results compare the qualitative harmonic analysis performed on a single table created by the intersection of two trajectories (i.e a sequence of states where states are intersection between geographical and residential states) to those obtained with an analysis of two data sets performed by the Multiple Factor Analysis (MFA).

References

Becue-Bertaut, M. & Pagès, J. (2008), Multiple factor analysis and clustering of a mixture of quantitative and Categorical frequency data, *Computational Statistics & Data Analysis*,**52**, 3255-3268

Deville, J.C & Saporta, G,. Analyse Harmonique Qualitative (1980), *Data Analysis and Informatics*, E.Diday ed., p375-389, North-Holland

Barbary O., Pinzon-Sarmiento L.M. (1998) L'analyse harmonique qualitative et son application à la typologie des trajectoires individuelles , Mathématiques et sciences humaines, **144**

Simultaneous Analysis of contingency tables drawn with telephone data registration from the National Telephone Service to Support Women Suffering Violence in Uruguay

Elena Ganón

Fundación Plenario de Mujeres del Uruguay (PLEMUU)

ganonelena@gmail.com

Keywords: Domestic Violence, Civil Society role, Gender mainstreaming, Correspondence Analysis

The National Telephone Service to Support Women Suffering Violence, phone 08004141, created in October 1992, is an example of successful and permanent link between the Non Governmental Organizations, Local Government and Telecommunication Companies. This paper analyzes the records issued from continued registration of phone calls, where several variables indicating the type of violence (origin (domestic/non-domestic), the form (physical/psychological and treated/executed)) and the social profile of the victim and the aggressor (age, education, occupation) among others are relieved. The separate and stacked correspondence analysis and simultaneous analysis of the contingency tables generated is made with a special focus on the temporal evolution in the period 2003 to 2009, and the characteristics of victim and aggressor in relation with the type of violence. The impact of changing technologies in the Service access and of the diffusion campaigns in the media is also considered.

References

Bécue_Bertaut, M., Pagès, J. (2004) *Multiple factor analysis for contingency tables*. In: Greenacre, M., Blasius, J. (Eds.), Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC, Boca Raton, FL. 300-326.

Escofier B., Pagès J. (1990) *Analyses factorielles simples et multiples, objectifs, méthodes et interpretation*. Dunod, Paris. 2nd Ed.

Greenacre M. (2007) *Correspondence Analysis in Practice*. Chapman & Hall/CRC, Boca Raton, FL. 2nd Ed.

Ganón E. (1995) El Servicio *en números. Evolución del número de llamadas. Estudio de caso: año 1994.* Publicado en: Carmen Tornaría (Ed) Un teléfono que da que hablar 414177. Fundación PLEMUU. Montevideo. Uruguay. 35-43, 63-91.

Ganón E. (2010) *Servicio Telefónico de Apoyo a la Mujer Victima de Violencia, Una experiencia uruguaya contada desde los números.* Presentado en: Congreso Internacional Las Políticas de Equidad de Género en Prospectiva: Nuevos escenarios, actores y articulaciones. Área Género, Sociedad y Políticas de FLACSO, 19 a 12 de noviembre de 2010, Buenos Aires, Argentina.

Lebart, L., Morineau, A. & Piron, M. (1997) *Statistique exploratoire multidimensionnelle*. Dunod, Paris 2nd Ed.

Walby, S.(2005) *Introduction: Comparative Gender Mainstreaming in the global era.* International Feminist Journal of Politics 7(4) December 2005, 453-470.

Zárraga A., Goitisolo B. (2006) Simultaneous analysis: A joint study of several contingency tables with different margins. In: Greenacre, M., Blasius, J. (Eds.) Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC, Boca Raton, FL. 327-350.

Zárraga A., Goitisolo B. (2009*) Simultaneous analysis and multiple factor analysis for contingency tables: Two methods for the joint study of contingency tables*. Computational Statistics and Data Analysis, 53(2009)3171-3182.

Semantics of Narrative in Collective, Distributed Problem-Solving Environments based on Correspondence Analysis and Hierarchical Clustering

Fionn Murtagh[1,2,*] **, Adam Ganz**[3] and Joe Reddington[1]
1. Department of Computer Science, Royal Holloway, University of London
2. Science Foundation Ireland, Dublin, Ireland
3. Department of Media Arts, Royal Holloway, University of London
* Contact author: fmurtagh@acm.org

Keywords: Multiple correspondence analysis, text analysis, contiguity-constrained hierarchical clustering, film, games

Our work has focused on support for film or television scriptwriting. Since this involves potentially varied story-lines, we note the implicit or latent support for interactivity. Furthermore the film, television, games, publishing and other sectors are converging, so that cross-over and re-use of one form of product in another of these sectors is ever more common. Technically our work has been largely based on all pairwise interrelationships that are used to reveal the semantics of the data, and the dynamics of the narrative revealed through change and anomaly on varying scales. The former, semantics, are operationalized through the Euclidean embedding provided by correspondence analysis. The latter, dynamics, are operationalized through an ultrametric embedding provided by a sequence- or temporal-constrained hierarchical clustering. We also discuss how our data analysis platform can support collective, distributed problem-solving.

References

F. Murtagh, A. Ganz and J. Reddington (2010). New methods of analysis and semantics in support of interactivity, *Entertainment Computing*, in press (Advance Access online).

F. Murtagh and A. Ganz (2010). Semantics from narrative, in S. Bolasco, I. Chiari and L. Giuliano, Eds., *Statistical Analysis of Textual Data, Proceedings of 10th International Conference JADT Journées d'Analyse Statistique des Données Textuelles*, 9–11 June 2010, Sapienza University of Rome, Vol. 1, pp. 443–453, LED Edizioni Universitarie di Lettere Economia Diritto, Milan, 2010.

F. Murtagh, A. Ganz and S. McKie (2009). The structure of narrative: the case of film scripts, *Pattern Recognition*, 42, 302–312, 2009.

F. Murtagh, A. Ganz, S. McKie, J. Mothe and K. Englmeier (2010). Tag clouds for displaying semantics: the case of filmscripts, *Information Visualization Journal*, 9, 253–262.

Cultural Distinctions: A Geometric Data Analysis.

Johs. Hjellbrekke & Olav Korsnes
Dep. of Sociology,
University of Bergen
Norway
johs.hjellbrekke@sos.uib.no

Inspired by Bourdieu's classic study "Distinction" (Bourdieu 1979) and by Le Roux and Rouanet's "Geometric Data Analysis" (Le Roux & Rouanet 2010), this paper offers an analysis of the structures of taste and cultural preferences in Norway, which often is preceived as an *egalitarian* society (Hjellbrekke & Korsnes 2006). The data originate from "The Culture and Media Survey 2008", distributed to a representative sample of Norwegians 18 yrs and older (N=1450). 44 questions on 6 different topics have been included in the analysis. These are variables on

- TV-preferences,
- participation in cultural activities
- music preferences
- newspaper and magazine readership
- interest in books and literature
- radio listening preferences

By way of specific multiple correspondence analysis, hierarchical cluster analysis, concentration and confidence ellipses and class specific analysis (see Le Roux & Rouanet 2010), a 3-dimensional space with 7 clusters is identified and examined in greater detail. Overall, the results contradict the claims by Chan & Goldthorpe regarding the social stratification of cultural preferences (2005, 2007a,b), but are more accordance with the analyses of Le Roux, Rouanet, Savage and Warde (2008) and Bennett & al. (2009) of the UK-case. Differences in cultural preferences and practices are still related to class inequalities, also in egalitarian societies.

Keywords: geometric data analysis, sociology, statistical inference in GDA, Class Specific Analysis (CSA)

References:

Bennett, Savage, Silva, Warde, Gayo-Cal & Wright (2009). *Culture, Class, Distinction.* London: Routledge

Bourdieu, P. (1984 [1979]): *Distinction. A Social Critique of the Judgment of Taste*. London: Routledge and Kegan Paul.

Chan, Tak Win & Goldthorpe, John H.(2005). "The Social Stratification of Theatre, Dance and Cinema Attendance." In *Cultural Trends* Vol. 14(3), No. 55, September 2005, pp. 193–212.

Chan, Tak Win & Goldthorpe, John H.(2007a). "Social Status and Newspaper Readership." In *American Journal of Sociology*, vol. 112, 4, pp. 1095-1134.

Evaluation of seminars by Correspondence Analysis and Related Methods

Françoise Bernard[1], Mireille Gettler Summa [2*], Bernard Goldfarb[2], Catherine Pardoux[2], Myriam Touati[2]

1. Institut FB – Paris France
2. Université Paris Dauphine - CEREMADE

* Contact author: summal@ceremade.dauphine.fr

Keywords: time series, textual data, mixed data, seminar evaluation, Correspondence Analysis, Clustering

We present a Correspondence Analysis and related methods approach which can be used as a tool in order to evaluate seminars.
We suggest some grids to collect both numerical and textual data for a single respondent: as seminars have some duration the collected data have an evolution and may thus be studied as time series in a mixed framework: continuous, categorical and textual. A case study dealing with sociology of education is used as a support for the presentation.

References

Abecassis P., Batufoulier P., Bilon I., gannon F., Martin B.(2007) Evolution of the French health system : a lexical analysis
http://economix.u-paris10.fr/docs/94/Athnes_-_Abecassis-Batifoulier-Bilon-Gannon-Martin.pdf
Bernard F. (2001), « La démarche. Autographie-Projets de vie ® avec les enseignants », *Journal français de psychiatrie, P.35-37*
Jacob S. et Ouvrard L. (2009), L'évaluation participative, avantages et difficultés d'une pratique innovante *Cahiers de la performance et de l'évaluation*, Québec, PerfEval, n° 1.
Lebart L., Salem A., (1994), *Statistique textuelle*, Dunod

Mapping a Citational Universe: A GDA of Literary Dissertation Bibliographies

Bo G Ekelund[1,2,*] , Mikael Börjesson[1,3]

1. Sociology of Education and Culture, Uppsala University
2. Department of English, Stockholm University
3. Second affiliation of author B

* Contact author: bo.ekelund@english.su.se

Keywords: Geometric Data Analysis, Literary studies, Bibliographies, reception of modern theory

In a study of the symbolic reproduction of non-Swedish works of literature, criticism and theory within the Swedish field of literary studies and criticism, the bibliographies of literary Ph.D. dissertations defended between 1980 and 2005 are analyzed in order to reveal the complex pattern of reception of non-domestic theory and criticism in the Swedish field of literary studies. For this analysis, the "citational universe" of the bibliographies is subjected to Geometric Data Analysis (Le Roux and Rouanet, 2004), and in particular specific Multiple Correspondence Analysis. Starting with a total of nearly two hundred thousand bibliographical posts from 680 dissertations, we narrowed down the data by first removing the primary sources, comprising roughly a fifth of all references. Of the fifty thousand individual authors cited among the secondary sources, we then selected those individuals who were cited at least five times and in at least two dissertations. It is these "frequently cited critics" and in particular the 2283 non-Swedish critics among them that were coded in order to bring out the space of critical choices made by Swedish doctoral students in this period.

Our analysis of this particular "citation culture" (cf. Wouters) gives us a preliminary view of the network of mediations, translations and intellectual flows that makes possible the production and reproduction of "Swedish" literary life. It also gives a view of the world system of literary theory, as seen from a semi-peripheral national field.

References

Le Roux, B and H Rouanet (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Kluwer (Dordrecht, Netherlands).

Wouters, P (1999). *The Citation Culture*. Diss. University of Amsterdam.

Nominal, Ordinal and Metric Variables in the "Social Space" – Using CatPCA to Examine Lifestyles and Regional Identities in a Medium-sized German City

Andreas Mühlichen[1,*]

1. Institute of Political Science and Sociology, University of Bonn
* Contact author: a.muehlichen@uni-bonn.de

Keywords: social space, CatPCA

Traditionally, to construct a "social space" in the tradition of Bourdieu multi-response-questions are employed yielding dichotomous data. CA or MCA are then used to construct the space and mostly a two-dimensional map is generated for the visualization. In our data set there is nominal, ordinal and metric data. To preserve the full information of the data but still be able to use a social space approach we use CatPCA instead of MCA. This approach allows us to include ordinal and metric variables on their higher measurement.

The data has been collected in Pulheim, a medium sized-German city of approximately 54,000 inhabitants, in co-operation with the Rhineland Regional Council (LVR). 382 members of different registered societies in Pulheim were interviewed using an online questionnaire. Among others, it entailed questions concerning lifestyle, commitment to club membership, and regional identity. To operationalize lifestyle parts of Pierre Bourdieu's questionnaire used in la distinction (1979) have been adapted to the German situation and thus characteristics of furniture and of clothing were measured as multi-response-questions. The questions concerning regional identity have been measured using the four-point Likert scales. Here the items are either used on an ordinal level or scales are constructed with the help of CatPCA. In the latter case the resulting object scores on each dimension have mean values of zero and standard deviations of one. The results are visualized by integrating biplots for ordinal and metric variables into a map of categorical variables.

References

Bourdieu, Pierre (1979). *La distinction: Critique sociale du jugement*, Editions de minuit.

Cross-over Methodologies: Correspondence analysis as a framework for mixed methods.

Jan Thorhauge Frederiksen, Lecturer, University College Zealand, Part-Timer Lecturer, Dep, of Psychology and Educational Science, Roskilde University, janjaf@ruc.dk

Keywords: Recruitment, Geometric data analysis, Correspondence analysis, Mixed Methods, Cultural Capital

The paper proposes and demonstrates the use of multiple correspondence analysis as a framework for embedding qualitative data in quantitative analyses. Whereas the position of mixed methods in general(e.g. Johnson 2004, Creswell & Clark 2006) retains the incongruence of such methods, the proposed cross-over methodology directly connects different forms of data at an empirical level, allowing for concurrent analysis of quantitative and qualitative aspects of a population. The data stems from a study(Frederiksen 2010) of how New Public management and service-oriented professionalizing of the public sector affect first the recruitment and second, the training of professionals. Through geometric data analysis and classification (Le Roux & Rouanet 2004, 2010), it is shown how recent changes in educational policies in Denmark have forced professional training sites into fierce competition, and radical expansion of their recruitment practice, by examining the educational and vocational trajectories of social educator students, showing how changes in recruitment leave professional training with a new student population. The study further embeds empirical data from qualitative interviews and fieldwork in the geometrical data analysis, allowing minute studies of the classroom practices and social biographies of students with different trajectories. The student practices and educational strategies can be sited in the quantitative data directly, and allows for a number of complex comparisons, demonstrating the feasibility and unique perspective of the cross-over methodology.

References

Creswell, J & Clark, V.(2006): Designing and Conducting Mixed Methods Research, Sage, London
Frederiksen, J.T.(2010): Between Practice and Profession, Roskilde University Press, Roskilde.
Johnson, R. B. (2004). Quantitative, Qualitative, and Mixed Research. EDUCATIONAL RESEARCHER October 2004 vol. 33 no. 7 14-26
Le Roux, B. & H. Rouanet (2004). Geometric Data Analysis. Dordrecht, Kluwer Academic Publishers.
Le Roux, B. & H. Rouanet (2010): Multiple Correspondence analysis. Sage, London.

History of canonical correspondence analysis (CCA) in ecology

Cajo J.F. ter Braak[1,*]

1. Biometris, Wageningen University and Research Centre, Box 100, 6700 AC Wageningen, the Netherlands
* Contact author: cajo.terbraakl@wur.nl

Keywords: correspondence analysis, history, external constraints, duality diagrams

In 1986, canonical correspondence analysis entered Ecology (ter Braak 1986) as a multivariate method to relate species abundance data to environment data from the same set of sites. The method was derived as an approximation to the maximum likelihood equation of a non-linear, unimodal latent variable model, and shortly thereafter (ter Braak 1987) as a method that provides the linear combination of predictors that best separates species niches. Independently, Chessel, Lebreton, Yoccoz and Sabatier (1987; 1988a; 1988b; 1989) invented the method as correspondence analysis variant of principal component analysis with instrumental variables (redundancy analysis) and as a generalization of linear discriminant analysis and dual scaling. Now, 25 years later, the founding paper is cited more than 2000 times. Here I reflect on the origin of the method, its uses in Ecology, the role of weighted averaging (principe barycentric), duality diagrams and on the assumptions/conditions under which the method works well.

References

Chessel D, Lebreton JB, Yoccoz N (1987). Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. *Revue Statistique Appliqué*, **35**, 55-72.

Lebreton JD, Chessel D, Prodon R, Yoccoz N (1988a) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Oecologia Generalis*, **9**, 53-67.

Lebreton JD, Chessel D, Richardot-Coulet M, Yoccoz N (1988b). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. *Acta Oecologia Generalis*, **9**, 137-151.

Sabatier R, Lebreton J-D, Chessel D (1989). Multivariate analysis of composition data accompanied by qualitative variables describing a structure. In: Coppi R, Bolasco S (eds) Multiway data tables. North-Holland, Amsterdam, pp 341-352.

ter Braak CJF (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**,1167-1179.

ter Braak CJF (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, **69**, 69-77.

Analyzing spatial multivariate structures

Stéphane Dray
Laboratoire Biométrie et Biologie Evolutive - CNRS UMR 5558
Univ. C. Bernard - LYON I
43 Bd 11 Novembre 1918
F-69622 VILLEURBANNE CEDEX, FRANCE
E-mail : stephane.dray@univ-lyon1.fr

Keywords: Moran's I, multivariate analysis, spatial autocorrelation, spatial weighting matrix.

Standard multivariate analysis methods aim to identify and summarize the main structures in large datasets containing the description of a number of observations by several variables. In many cases, spatial information is also available for each observation, so that a map can be associated to the multivariate dataset. Two main objectives are relevant in the analysis of spatial multivariate data: summarizing covariation structures and identifying spatial patterns. In practice, achieving both goals simultaneously is a statistical challenge, and a range of methods have been developed that offer trade-offs between these two objectives. In an applied context, this methodological question has been and remains a major issue in community ecology, where species assemblages (i.e., covariation between species abundances) are often driven by spatial processes (and thus exhibit spatial patterns).

I review a variety of methods developed during the last decades to investigate multivariate spatial patterns. I present different ways of incorporating spatial constraints in multivariate analysis (e.g., geographical distance, spatial partition, polynomial of geographical coordinates, spatial graph) and discuss the properties of these different approaches both from a practical and theoretical viewpoint.

The Poisson trick for matched tables: a case for putting the fish in the bowl

Simplice Dossou-Gbété[1] **, Antoine de Falguerolles**[2,*]
1. Université de Pau et des Pays de l'Adour
2. Université Paul Sabatier (Toulouse III)
* Contact author: falguero@cict.fr

Keywords: biplot visualization, matched tables, Poisson trick, correspondence analysis

Putting the fish in the bowl or the bird in the cage are simple experiments in retinal convergence. The Poisson trick is used in the context of categorical data analysis to fit binomial (multinomial) regression models by assuming independent Poisson distributions for the counts.

This papers addresses the topic of visualization of the interactions in (a pair of) matched tables. The situation of matched tables is exemplified by the data on suicide in Germany where the counts are classified by two factors A and B (class age and method of suicide) observed on two subpopulations, a factor R (gender) with two levels. The statistical analysis focuses on interactions between factors A, B and R.

Broadly speaking, analyses are two-steps: the first consists in some form of pre-processing of the data, while the second focuses on the visualizations of some restricted high order interactions. The contingency table under consideration being denoted by y^{ABR}, examples for the first step are constructing a common table and specific tables, or coercing into matrix form the residuals of some log-linear model, namely [AR][B], or [A][BR], or [AB][AR][BR]. The second step consists in analyzing the tables obtained in the first step by correspondence analysis (generalized singular values) or some variant (generalized bilinear models).

Consider now the relative proportions obtained by dividing the counts in one table, say y^{ABR}_{ab2}, by the corresponding counts in the other, y^{ABR}_{ab1}. This defines a two-way matrix of empirical odds ratios with general term

$$Z^{AB}_{ab} = \frac{Y^{ABR}_{ab2}}{Y^{ABR}_{ab1}} = \frac{Y^{ABR}_{ab2}}{(Y^{AB}_{ab} - Y^{ABR}_{ab2})}.$$

This is the framework for binomial regression $\mathcal{B}(\pi^{AB}, y^{AB})$ where the unknown probabilities π^{AB} possibly depend on the explanatory variables A and B, and where the known parameters are given by $y^{AB} = y^{AB(R=1)} + y^{AB(R=2)}$. In other terms, the matched fish and bowl (or bird and cage) are now superimposed and the Poisson trick tells how a model for the binomial regression with logit link corresponds to a log-linear model for the initial three way table. Of special interest are the following situations:

Binomial regressions $\mathcal{B}(\pi^{AB}, y^{AB})$	Log-linear analyses of table y^{ABR}
additivity of effects $[A][B]$	all two-way $[AR][BR][AB]$
saturated model $[AB]$	saturated model $[ABR]$
reduced rank $[AB]$	reduced rank interactions $[ABR]$

Note that the Poisson trick extends two more than two occasions (tables) although the implementation is less straightforward.

Special attention is given to biplot visualisations for the restricted interactions. The visual effects due to open options in the selection of variance function and link function, or in the choice of identification constraints are investigated.

References

Dossou-Gbété, S., and Grorud, A., (2002). Biplots for matched two-way tables *Annales de la faculté des sciences de Toulouse* Sér. 6, 11(4), 469-483.

Greenacre, M. J., (2003). Singular value decomposition of matched tables. *Journal of applied statistics*, **30**, 1101–1113.

van der Heijden, P. G. M., & Worsley, K. (1988). Comments on correspondence analysis used complementary to loglinear analysis. *Psychometrika*, **53(2)**, 287-291.

The Aggregate Prediction Index and Non-Symmetrical Correspondence Analysis of Aggregate Data: The 2×2 Table

Eric J. Beh[1*] **, Rosaria Lombardo**[2]
1. School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia
2. Department of Strategy and Quantitative Methods, Second University of Naples, Gran Priorato di Malta, 81043 Capua (CE), Italy
* Contact author: e.beh@uws.edu.au

Keywords: Predictability; Contingency Table; Ecological Inference; Profile Coordinates; Graphical Display.

In ecological inference the analysis of a single 2×2 contingency table poses one of the most enduring and intriguing of questions "if only the marginal information is known, what can it tell us about the association between the categorical variables?". Furthermore, if such an association exists, "how are we able to graphically depict this association?". Considerable literature is available concerning the quantification of association coefficients for 2×2 tables (see, for example, Janson & Vegelius, 1981; Baulieu,1989; Warrens, 2008).

Recently, addressing this issue has lead to the development of the aggregate association index (AAI, Beh, 2008) and of the correspondence analysis of aggregate data (Beh, 2010). Here we expand on this work by studying the case when two dichotomous variables have an *a priori* established role such that one is a predictor variable and the other is a response variable. When only the marginal information of a single 2×2 contingency table is available, our aim is to investigate the strength of prediction between the categorical variables. We will propose an aggregate prediction index (API) (akin to Beh's (2010) AAI) when considering the ecological regression (Goodman, 1959) and the Goodman-Kruskal tau index (1954).

Furthermore, since the Goodman-Kruskal tau index lies at the heart of non-symmetrical correspondence analysis (NSCA; D'Ambra and Lauro, 1989; Lauro and D'Ambra, 1984) we discuss the applicability of the API to NSCA to provide a graphical depiction of the predictive relationship of the rows, given the columns, when the joint cell frequencies are not available. We will present a comparison between classical plots and biplot graphical displays (Kroonenberg and Lombardo, 1999).

References

Baulieu, F.B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 233 - 246.

Beh, E. J. (2008). Correspondence analysis of aggregate data: The 2×2 table. *Journal of Statistical Planning and Inference*, **138**, 2941 - 2952.

Beh, E. J. (2010). The aggregate association index. *Computational Statistics & Data Analysis*, **54**, 1570 - 1580.

D'Ambra, L. and Lauro, N. C. (1989). Non-symmetrical correspondence analysis for three-way contingency table. In *Multiway Data Analysis*, (R. Coppi and S. Bolasco Eds.), Amsterdam: Elsevier, 301 – 315.

Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, **49**, 732–764.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *The American Journal of Sociology*, **64**, 610-625.

Janson, S., and Vegelius, J. (1981). Measures of ecological association. *Oecologia*, **49**, 371-376.

Lauro, N.C. and D'Ambra, L. (1984). L'Analyse non symetrique des Correspondences. In *Data Analysis and Informatics III* (eds Diday E. et al.), 433–446. Amsterdam: Elsevier.

Kroonenberg, P., and Lombardo, R., (1999). Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, **34**, 367-397.

Warrens, M. J. (2008). On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, **73**, 777–789.

Regularized generalized canonical correlation analysis

Arthur Tenenhaus[1] & Michel Tenenhaus[2*]

1. SUPELEC
2. HEC Paris

* Contact author: tenenhaus@hec.fr

Keywords: Generalized canonical correlation analysis, Multi-block data analysis, PLS path modeling, Regularized canonical correlation analysis

Regularized generalized canonical correlation analysis (RGCCA) is a generalization of regularized canonical correlation analysis to three or more sets of variables. It constitutes a general framework for many multi-block data analysis methods. It combines the power of multi-block data analysis methods (maximization of well identified criteria) and the flexibility of PLS path modeling (the researcher decides which blocks are connected and which are not). Searching for a fixed point of the stationary equations related to RGCCA, a new monotone convergent algorithm, very similar to the PLS algorithm proposed by Herman Wold, is obtained. Finally, a practical example is discussed.

Reference

A. Tenenhaus and M. Tenenhaus (2011), Regularized generalized canonical correlation analysis. *Psychometrika*.

Visual Displays. Some evidence through artificial and real data

K. Fernández-Aguirre[1,*], M. A. Garín-Martín[1], J. I. Modroño-Herrán[1]

1. University of the Basque Country (UPV/EHU), Bilbao, Spain

* Contact author: karmele.fernandez@ehu.es

Keywords: Visual Displays, Principal Component Analysis, Correspondence Analysis, Clustering

In recent years, the main objective for most practitioners is to identify interesting structures in the data sets, such as clusters of observations, or relationships among the variables. Principal axes methods such as Principal Component Analysis (PCA) and Correspondence Analysis (CA) are useful for the identification of structures in the data through interesting graphical visualizations. However, some kinds of data sets could be treated alternatively by PCA or CA.

In the literature, PCA often appears as the method best suited to the analysis of quantitative variables measured on different observations or individuals, two-way Correspondence Analysis (CA) to the analysis of contingency tables that cross two categorical variables and Multiple Correspondence Analysis (MCA), or any of his variants, as the method of choice for the analysis of a table of categorical variables coded in either a disjunctive complete form (indicator matrix) or a table of multiple correspondences crossing more than two categorical variables (Burt matrix). A comparative analysis of these possibilities can be found in Tenenhaus and Young (1985).

These methods are applied in almost all areas of knowledge where predilection for each of them is variable. In certain areas in particular, it is still frequent the treatment of categorical variables as if they were continuous, due to the great influence of the classic school that dates back to the beginning of the 20th century, see Gifi (1990), Chapter 1. A recent reference providing the state of the art of the data analysis of qualitative and categorical variables can be found in Greenacre and Blasius (2006). The book offers an exhaustive view of the most different aproaches of CA and MCA.

As possible examples of choices of the analyses cited above, one can consider, for instance, a data matrix that measures the number of employees in different economic sectors for the countries of the European Union. Such matrix can be considered to be a matrix of quantitative variables and a PCA be applied on it, or a contingency table and a two-way CA the analysis to be performed on it. Another example would be a matrix containing the answers of a survey in an ordinal scale, which can be treated by means of a PCA, a Categorical PCA or a MCA, though the latter is not always admitted. These possibilities of application take to comparable but different results, depending on the characteristics and the properties of each method.

Our emphasis in the following discussion is on methods, such as PCA and CA, and visual displays. This paper has two parts. In the first part, we analitically study the case of a binary matrix M associated to a symmetric graph G (Octagon), also valid for the cases of high dimensionality graphs, showing the superiority of CA for the reconstitution and visualization of such symmetric graphs over the visualization obtained with PCA, see Lebart et al. (1998), pp. 63-69. In the second part, we present a case of actual data on the distribution of employees in different economic sectors for the countries of the European Union analyzed by means of PCA (PCA with transformation of variables) and two-way CA. The results are complemented with cluster analysis. In this way, we can illustrate clearly the implications, for a potential user, of the selection of a method with respect to an alternative one from an applied point of view, and the advantages or disadvantages of such methods.

References

A. Gifi,(1990). *Non Linear Multivariate Analysis*, Wiley, Chichester.

M. J. Greenacre, & J. Blasius, (eds.)(2006). *Multiple Correspondence Analysis and Related Methods.* Chapman & Hall/CRC, London.

L. Lebart, A. Salem, & L. Berry (1998) *Exploring Textual Data*, Kluwer Academic Publishers, New York.

M. Tenenhaus & F. W. Young (1985). *An analysis and synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other methods for quantifying categorical multivariate data*, Psychometrika, 21, pp. 91-119.

Hierarchical Clustering on Special Manifolds

Angelos Markos[1*], George Menexes[2]

1. Laboratory of Mathematics & Computer Science, School of Primary Education, Democritus University of Thrace, Greece
2. Laboratory of Agronomy, School of Agriculture, Aristotle University of Thessaloniki, Greece

* Contact author: amarkos@eled.duth.gr

Keywords: Geometric data analysis, Hierarchical clustering, Matrix concordance, Riemannian manifolds

In this work, we address the problem of comparing a set of vectors to other sets of vectors, which naturally corresponds to a clustering problem on spaces of orthogonal linear projections. Such data arise in earth and biological sciences, medicine, computer vision and signal processing. In this context, we review measures for calculating distances between orthonormal matrices and between equivalence classes of matrices that span the same subspace. All distances can be represented with principal angles and their relationships with well established similarity criteria, such as the RV coefficient, are also considered. We adopt two notions of the mean or centroid of subspaces, each associated with a different distance metric: the Karcher mean, which minimizes the sum of squared geodesic distances and a Procrustes mean relying on the embedding of a manifold in the ambient Euclidean space. By exploiting the differential geometry of special Riemannian manifolds, we introduce some hierarchical clustering methods to efficiently group sets of orthonormal matrices. The proposed methods are demonstrated using synthetic and real data.

References

Absil, P.A., Mahony, R. and Sepulchre, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press.

Chikuse, Y. (2003). *Statistics on special manifolds*, Lecture Notes in Statistics, vol. 174, Springer, New York.

Edelman, A., Arias, T., & Smith, S. (1998). The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, **20**(2), 303–353.

Golub, G.H. & Van Loan, C.F. (1996). *Matrix Computations*, Johns Hopkins University Press, Baltimore.

Robert, P. & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied Statistics*, **25**, 257–265.

CD-clustering

Cristina Tortora [1,2,*], **Francesco Palumbo**[3], **Mireille Gettler Summa**[2]
1. Dip. di Matematica e Statistica, Univ. di Napoli Federico II
2. CEREMADE, Univ. Paris Dauphine
3. Dip. di Scienze Relazionali "G. Iacono", Univ. di Napoli Federico II
* Contact author: cristina.tortora@unina.it

Keywords: Categorical data, Distance Clustering, Multiple Correspondence Analysis.

CD-clustering is an adaptation of probabilistic D-clustering (Ben-Israel, A. & Iyigun, C. 2008) to the case of categorical data. Clustering of categorical data presents well known issues: categorical data can be combined in order to determine a limited subspace of the global data space. Indeed these type of data are thus characterized by non-linear associations that often remain invisible to classical clustering techniques.

Probabilistic D-clustering is an iterative method for probabilistic clustering of data. Dealing with categorical data we can not use an euclidean metric and this method can not be applied.

We propose to combine two approaches in order to obtain enhanced results. Categorical data quantification step is introduced in the D-Clustering procedure in order to adapt the algorithm to categorical data. To do this it is necessary that quantification method and clustering method optimize the same criteria. This type of two steps strategy, based on a quantification and a classification step, is wildly used (Arabie, P & alt. 1996). Starting from this approach a lots of iterative techniques are developed, they iterate the two steps until convergence.

Probabilistic D-clustering can be summarized as follow: given some random centers, the probability of any point to belong to each class is assumed inversely proportional to the distance from the center of the cluster. At each iteration centers are computed as a convex combination of the points. This method assumes that the product between the distance of each point from a center of each clusters and its probability to belong to this cluster is a constant $D(x)$ depending on the point x, called joint distance function (JDF). JDF is a measure of the distance of x from all cluster centers so it measures the classificability of the point x. If it is zero, the point coincides with one of the cluster centers, in this case the point belongs to the class with probability 1. If all the distances between the point and the k centers of the classes are equal, in particular equal to d, $D(X) = d/k$ and all the probabilities to belong to each class are the same $p(x) = 1/K$. Consequently the objective is to minimize the sum of the JDF.

We propose an iterative method in order to adapt probabilistic D-clustering to categorical data. The first problem with categorical data is that the usually adopted complete binary coding leads to very sparse and large binary data matrices. In order to solve this problem and to quantify the original dataset we apply a MCA on raw data matrix. It permits to preserve the non-linear association structure and to reduce the number of variables (Saporta 1990). The method can be summarized as follow:

- MCA on the row data matrix
- probabilistic D-clustering on the first factorial axes
- projection of data in the space that optimize the same criteria of the probabilistic D-clustering

We iterate the second and the third steps until convergence. Empirical trials have demonstrated procedure converges. Probabilistic D-clustering, as other classical methods, works well when dealing with class of hyperspherical form. It cannot find clusters of arbitrary form. Projecting the points on a new space can help to simplify the structure of this type of cluster. The iterative method allows to find clusters not hyperspherical or even nested. The quantification of data on few factorial axes bring us to visualize the data and it can be an important advantage in the interpretation of results.

References

Arabie, P & Hubert, L. J. & De Soete (1996). *Clustering and Classification*, Word Scientific Publ., River Edge, NJ.

Ben-Israel, A. & Iyigun, C. (2008). Probabilistic D-clustering. *Journal of Classification*, **25**, 5–26.

Saporta, G. (1990). Simultaneous analysis of qualitative and quantitative data. *Atti* 35° *Riunione Scientifica della Societá italiana di Statistica*, CEDAM, 63–72.

Urban Aboriginal lifestyles in Brisbane: mapping vertical and lateral stratification of opportunity for marginalised groups

Robert Funnell, Faculty of Education – Griffith University, Nathan Q 4111 Australia (Email r.funnell@griffith.edu.au)

Keywords: space, lifestyles, urban sociology

Abstract

This paper seeks to explain how correspondence analysis can be used to enhance research about 'urban Aboriginal' populations in Australia. Since the late 1960s, when Aboriginal peoples were first included in the national census, it has been difficult to make clear comparisons across the lifestyles of Aboriginals, other 'ethnic' groups and 'other Australians'. Here the census remains an imprecise instrument; it provides a vertical scale from which Aboriginal policy decisions are made. But this census information cannot be realistically disaggregated to other dimensions of stratification that is, to the particular urban living conditions in which Aboriginal people enter into social relations with others. It is argued that surveys outside of the census are required from which it would be possible to describe the spread of Aboriginal and non-Aboriginal lifestyles across various social groupings or strata. The author has conducted ethnographic research and later administered a survey with a cross-section of five hundred residents in the outer suburbs of Brisbane. The paper focuses on a preliminary discussion of the extent to which relations between groupings can be shown through the use of correspondence analysis. Conclusions are made about the potential of vertical and lateral scales to extend the census in understanding differences in the lifestyles of marginalised groups.

Social and spatial structures in an urban environment

Lennart Rosenlund
University of Stavanger

Keywords: multiple correspondence analysis, principal component analysis, cluster analysis, urban differentiation, social space, space of lifestyles, volume and composition of capital, Bourdieu

Abstract

This contribution is founded on findings from a thorough, empirical study of a specific urban community that has undergone a profound and rapid process of social change, Stavanger, the oil capital of Norway. It begins with reflections on how to construct a social structure that synchronically and diachronically is able to catch the most potent mechanisms of social division and of relations of domination in contemporary society. Pierre Bourdieu's conception of "the space of social positions" is then introduced as a useful device for such a venture. This conception postulate that processes of social differentiation should be conceived as multidimensional phenomena and that the distributions of economic and cultural capital are pivotal for their understanding. In doing so a recent survey of lifestyles among the citizens is exploited.

It then proceeds by examining a version of Bourdieu's second space construct that of the "space of lifestyles". This is a representation of divisions and contradiction within a universe of finely differentiated set of beliefs, practices, symbols and strategies, both conscious and unconscious, all products of differentiated habituses. What emerges are relations of homology; the universe of basic conditions of existence (the space of social positions) and the universe of beliefs, practices and symbols (the space of lifestyles) are governed by the very same principles of differentiation: volume and compositions of capital. Finally, within the infinite space of lifestyles it is possible to establish a particular symbolic space consisting of imageries of the various residential areas, which is structured by the same set of principles (volum and composition of capital). The inhabitants have "practical knowledge" about their community that has been developed being a citizen of it. They "know" where they would fit in and where they don't. They tend to favour areas where their own sorts (social positions and lifestyle configurations) are prevalent and they tend to reject those where they are few and where they would have been excluded. Multiple correspondence analysis (MCA) is the analytic device in this venture.

Then these analyses are supplemented by the exploration of data on living condition in the city produced by the municipality. The city has been divided into 68 homogenized zones with approximately equal number of inhabitants in each. The database contains vital statistics of each of these zones. The analysis has been undertaken with GDA (APC and cluster analysis) and the presentation of the results is aided by the help of maps and photographs. The results indicate that social agents in dominant positions in the space tend to favour "high status", high price areas, those in dominated positions favour dominated areas (bad reputation, bad infrastructure etc.). Groups whose capital accounts are dominated by cultural capital, favour areas that are being gentrified, or have potentials of being so, while those whose capital assets are dominated by economic capital prefer the suburban areas in the periphery of the city. Seen in this way the spatial organization of the community is not only a physical reflection of the major forces of social differentiation, but it becomes a force of its own in the reproduction of relations of domination and inequality.

Latest methodological breakthroughs in geometric data analysis of cultural practices.

Philippe BONNET[1],* and Frédéric LEBARON[2]

Keywords: Geometric Data Analysis, Class Specific Analysis, Sociology, Structural Homologies.

Our proposal is to make an assessment of the latest breakthroughs in geometric data analysis. This will be illustrated with cultural practices data from the permanent INSEE survey (2003) about cultural and sport participation. The aim is to prolong the theoretical and methodological approach Pierre Bourdieu and Monique de Saint-Martin initiated in "L'anatomie du goût". This approach can be enriched with the new possibilities of geometric data analysis.

Different kinds of problems will be examined at different steps of analysis:

- preparation of the data table: choose active individuals and active variables and encode categories;
- choose the method (MCA, specific MCA);
- after interpretation of axes in the cloud of categories, inspecting and dressing up the cloud of individuals;
- supplementary elements: individuals and variables;
- deep investigation of the cloud of individuals (structuring factors, concentration ellipses, between-within variance,…);
- class specific analysis to examine a subcloud of individuals (the young, the working class,…);
- statistical inference.

These problems will be presented and illustrated within the analysis of cultural practices and lifestyles data.

References

Bourdieu, P. (1979). *La distinction. Critique sociale du jugement.* Paris: Minuit.

Le Roux, B. & Rouanet, H. (2010). *Multiple Correspondence Analysis* (QASS Series). Thousand Oaks,CA: Sage.

Rouanet, H., Ackermann, W., Le Roux, B. (2000). The geometric analysis of questionnaires: The lesson of Bourdieu's La Distinction. *Bulletin de Méthodologie Sociologique, 65,* 5-18.

[1] Laboratoire de Psychologie et Neuropsychologie Cognitive (LPNCog), FRE 3292, CNRS et Paris-Descartes.
[2] Centre Universitaire de Recherche sur l'Action Publique et le Politique (CURAPP), UMR 6054, Université de Picardie – Jules Verne et CNRS.
* Contact author : philippe.bonnet@parisdescartes.fr

The Swedish Social Space of 1990. Investigating its Structure and History

Mikael Börjesson[1*] , Andreas Melldahl[1]

1. Sociology of Education and Culture (SEC), Uppsala University
* Contact author: mikael.borjesson@edu.uu.se

Keywords: Geometric Data Analysis, specific Multiple Correspondence Analysis, Bourdieu, Social Space, Sweden.

As is well known, Pierre Bourdieu operates in La Distinction (1979) with three dimensions in the analysis of the structure of the social space: volume and structure of capital and the changes of the assets over time. In this paper we make use of these three dimensions in an investigation of the Swedish social space. Since we have access to the rich census material on the whole Swedish population, we use this to construct the space of social positions in 1990. This is done by the means of Geometric Data Analysis (GDA), in particular specific Multiple Correspondence Analysis (MCA) (Le Roux & Rouanet 2004). As active variables we will employ information on various types of income and housing, level of education and field of study, place of residence, sector of the labor market, time devoted to work and marital status.

While the volume and the structure of the capital construct a two-dimensional hierarchical space, the third dimension involves "a balance-sheet of former struggles." The space is—in other words— structured by its history, the collective trajectories of the social groups change over time and all strategies of reproduction are directed towards the future. This dimension is, however, often developed mainly on a theoretical level—where time is inscribed as the third dimension of the space—but less often illustrated and investigated empirically. In this paper, we investigate the possibilities to study this dimension by use of earlier census material. A useful feature of the Swedish official statistics is the individual identification number. This makes it possible to link individuals and generations from various data registries to each other—that is, it is possible to discern the occupational position a 50 year old physician in 1990 had twenty years earlier, or to link an individual's position in 1990 to the parents' positions in 1970.

We explore the possibilities to in this way introduce the history of the Swedish space in the analysis of its structure in two ways. We examine on the one hand the relation between the social position of all 30-year-olds in 1990 and their parents' positions in 1970 and on the other the changes of the social positions over time, their numerical development and their material and symbolic standings.

References

Bourdieu (1979). *La distinction. Critique sociale du jugement*, Paris: Minuit.

Le Roux, B. & Rouanet, H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Dordrecht, Boston, London: Kluwer Academic Publishers.

Out-of-Study Practices and Symbolic Capital Among Swedish Students in Higher Education

Ida Lidegran, Uppsala University, & Mikael Palme

Stockholm University
Campus Konradsberg
S-106 91 Stockholm

Keywords: specific MCA, sociology of education, higher education, symbolic capital, cultural capital)

Previous research indicates that the social structure of Swedish higher education – in spite of other differences – to a large degree mirrors the oppositions exposed in French higher education by Bourdieu and others. While elite institutions oppose popular ones, the former are polarized along a cultural-economic dimension (see for example Broady, D. & Palme, M., "Le champ des institutions de l'éducation supérieur en Suède", i Monique de Saint Martin (ed): *Les systèmes de l'enseignement supérieur et la formation des cadres dirigeants*, Centre de sociologie européenne, Paris, 1992). Using data from a student questionnaire (n=2500) collected in 2004-06, this paper further explores social and cultural differences among Swedish university students. With a focus on students pertaining to high-positioned institutions in the field of higher education in the Stockholm-Uppsala region, the article sets out to examine differences as regards students' out-of-study practices, as well as related beliefs and attitudes. Using specific MCA, the analysis unveils major oppositions between students who involve in intense and extended cultural activities and those characterized by abstention from such activities, between students who valorize expensive, body-oriented activities and those who rather opt for low-cost activities, and between students who display interest in traditional high-culture and those who prefer a less traditional and more youth-oriented culture. Using study program adherence, along with social origin, as structuring factors, the analysis gives a picture of the relevance of both study orientation and parental background for differences related to students' out-of-study practices. The patterns uncovered by the specific MCA are interpreted as an expression of differences related to investments in competing symbolic values, i.e. in forms of symbolic capital that to a large degree oppose each other, among students at elite higher education institutions oriented towards careers in social fields with differing mechanisms of recognition and cooptation.

Not so trustful after all? A study of trust, tolerance and solidarity in Denmark

Morten Frederiksen
* Contact author: mf@soc.ku.dk

This paper presents a mixed methods study of social trust and the way trust is part of the dispositional outlook formed by positions in social space. The analysis is based on the Danish wave of the European Value Study 2008 in combination with qualitative research interviews concerning experiences and dispositions of trust. Social positions based on cultural, social and economical capital is projected in a space of dispositions using sMCA. Interview participants are also projected into this space as supplementary individuals enhancing the interpretation of the modes of trust and the dispositional sets trust is part of. The paper argues for a strong relation between dispositions of trust, solidarity and tolerance in different constellations depending on position. Further more it is argued that dispositions of trust are intertwined with self-perception and feelings of empowerment associated with the experience of specific social positions. The expanse of social space to which one extends trust as disposition is homologous to the level of domination associated with the social position one inhabits. Euclidian classification is applied to study the divergence and convergence of studied positions in social space.

History of Nonlinear Principal Component Analysis

Jan de Leeuw
UCLA
http://www.cuddyvalley.org/

Keywords: Multiple Correspondence Analysis, Nonlinear Principal Component Analysis

Multiple Correspondence Analysis (MCA) is discussed as a form of Nonlinear Principal Component Analysis (NLPCA). It is compared with other forms of NLPCA that have been proposed over the years: Shepard-Kruskal- Breiman-Friedman-Gifi PCA with optimal scaling, aspect analysis of correlations, Guttman's MSA, Logit/Probit PCA of binary data, and Logistic Homogeneity Analysis.

Past, Present, and Future of Multidimensional Scaling

Patrick J.F. Groenen[1,*]
1. Econometric Institute, Erasmus University Rotterdam
* Contact author: groenen@ese.eur.nl

Keywords: Multidimensional scaling, Majorization, Visualization

The technique of multidimensional scaling (MDS) has established itself as a very useful tool for statisticians and applied researchers. Its success is due to the simplicity of the presentation of the results. These are often represented in two dimensional map with objects (attributes, stimuli, respondents) represented as points such that those are near to each other are similar and those far apart are different. Over the last 50 years, multidimensional scaling has become one of a standard technique in multivariate analysis.

In this paper, we pay tribute to several important developers of MDS and give a subjective overview of milestones in MDS developments. We also discuss the present situation of MDS and a brief outlook on its future.

Index

Asan Zerrin, 30

Bastos Ronaldo, 35, 55
Beh Eric, 17, 76
Ben Ammou Samir, 32
Benlagha Noureddine, 32
Bennani Dosse Mohammed, 48
Bernard Françoise, 69
Bienaise Solène, 53
Blasius Jörg, 54
Bonnet Philippe, 83
Bonvalet Catherine, 65
Bougeard Stéphanie, 21
Buche Marianne, 42, 52
Bécue-Bertaut Mónica, 46, 50
Bénasséni Jacques, 48
Böcük Harun, 30
Börjesson Mikael, 70, 84

Cadoret Marine, 42, 43, 52
Cadot Martine, 49
Camminatiello Ida, 17
Cariou Véronique, 44
Cazes Pierre, 28
Chavent Marie, 18
Choulakian Vartan, 16
Corcoran Paul, 58
Courcoux Philippe, 44
Csernel Marc, 23

D'Ambra Luigi, 17
de Leeuw Jan, 87
De Rooij Mark, 61
De Tibeiro Jules, 16
de Tibeiro Jules, 36
Dossou-Gbété Simplice, 38, 75
Dray Stéphane, 74
Durucasu Hasan, 31
Dziechciarz Jozef, 25

Ekelund Bo, 70

Fablet Christelle, 21
Falguerolles Antoine de, 75
Fernández-Aguirre Karmele, 78
Fersi Kmar, 32
Firth David, 60
Fitzgerald Tony, 58
Frederiksen Jan Thorhauge, 72
Frederiksen Morten, 86
Friendly Michael, 60
Funnell Robert, 81

Ganz Adam, 67
Ganón Elena, 66
Garnier Bénédicte, 65
Garín-Martín Maria Araceli, 78
Gettler Summa Mireille, 80
Goitisolo Beatriz, 56
Goldfarb Bernard, 69
Gower John, 59
Graffelman Jan, 62
Grannell Andrew, 58
Greenacre Michael, 15, 33
Groenen Patrick, 88
Grzeskowiak Alicja, 25

Hjellbrekke Johs., 68
Hornbostel Stefan, 51
Husson François, 18

Ican Özgür, 31
Iodice D'Enza Alfonso, 39

Josse Julie, 18

Karim Jahanvash, 24
Korneliussen Tor, 26
Korsnes Olav, 68
Kostov Belchin Adriyanov, 50
Kroonenberg Pieter, 47
Kuhnt Sonja, 19

Langovaya Anna, 19
Le Pouliquen Marc, 23
Le Roux Brigitte, 14, 53
Lebaron Frédéric, 83
Lebart Ludovic, 27
Lelu Alain, 49
Lidegran Ida, 85
Liquet Benoît, 18
Lombardo Rosaria, 76
Lubbe Sugnet, 57
Lê Sébastien, 22, 42, 43, 45, 52

Mair Patrick, 37
Markos Angelos, 79
Marty Christoph, 51
Melldahl Andreas, 84
Menexes George, 79
Modroño-Herrán Juan Ignacio, 78
Morand Elisabeth, 65
Morin Annie, 41, 50
Murdoch Duncan, 36
Murtagh Fionn, 67
Mühlichen Andreas, 71

Nenadic Oleg, 33
Niel le Roux, 57

Pagès Jérôme, 22, 43, 45
Palacios Fenech Javier, 34
Palme Mikael, 85
Palumbo Francesco, 39, 80
Pardoux Catherine, 69
Pham Nguyen-Khang, 41

Qannari El Mostafa, 21
Qannari Mostafa, 44

Reddington Joe, 67
Rosenlund Lennart, 82
Roux Maurice, 29

Saporta Gilbert, 20
Silal Sheetal, 57
Souza Augusto, 55
Souza Marcio, 35
Stanimir Agnieszka, 25
Summa-Getler Mireille, 69
Sund Reijo, 40
Séguéla Julie, 20

Tenenhaus Arthur, 77
Tenenhaus Michel, 77
ter Braak Cajo, 73
Tortora Cristina, 80
Touati Myriam, 69
Turner Heather, 60
Türe Cengiz, 30

Vehkalahti Kimmo, 40
Verbanck Marie, 22
Vicente-Villardon Jose Luis, 63
Vieira Marcel, 35, 55
Vines Karen, 64
Volpato Richard, 13

Weisz Robert, 24
Worch Thierry, 45
Wurzer Marcus, 37

Zeileis Achim, 60
Zárraga Amaya, 56