IWCA

First International Workshop of CARME

CARME in ASSOS

Correspondence Analysis and Related Methods

Assos, Turkey

Saturday 1st to Tuesday 4th October 2011

IWCA

Welcome to the International Workshop of CARME – CARME in ASSOS, the first of what we hope will become a regular event in the calendar of statisticians and applied researchers interested in multivariate data analysis and visualization. The event is an offshoot of the quadriennial CARME conferences (Correspondence Analysis and Related Methods) that started in 1991 at the *Zentralarchiv für Empirische Sozialforschung* in Cologne, Germany. The latest CARME conference, celebrating 50 years of correspondence analysis (CA), took place in Rennes, France, in February this year, so we are still in that celebration year! Jean-Paul Benzécri originally defined the method, as we know it in its most popular form today, at the start of the 1960s in Rennes.

We are in the area of ***Assos***, which is more well-known by its Greek name than the Turkish name of *Behramkale*. Assos was founded about 2800 years ago by colonists from the town of *Mythimna* (modern-day Molivos) on the island of Mytilini (Lesbos). On top of a hill surrounded by olive groves are the ruins of the Doric-style *Temple of Athena* (530 BC) surrounded by crumbling city walls and an ancient necropolis (cemetery). Nearby is the 14th-century Ottoman *Murad Hüdavendigar* mosque. The hill offers spectacular views of the Aegean Sea and the nearby Greek island of Lesbos. Down the steep seaward side of the hill is the *iskele* (wharf, or harbour) of Assos, with old stone houses now serving as inns, hotels and restaurants.

From 341 BC ***Aristotle***, aged 37, lived here in Assos for three years and was married to Pythias, the niece of Hermias, the king who ruled the area. While in Assos, Aristotle became the leader of a group of philosophers and scientists who observed and discussed the anatomy, structure and classification of various plants, animals and insects. In this way Assos can be considered as the ancient source of the science of classification. In 344 BC, the Persians attacked Assos and killed Hermias. Aristotle escaped and journeyed to Macedonia, stopping on the way in Lesbos for a year to continue his study of biology, with his vegetarian pupil Theophrastus who was studying the plant kingdom.

We have planned a meeting here in Assos where you can relax and meet colleagues from many different countries, hear some interesting new presentations, enjoy some wonderful Turkish food, and go on several interesting excursions in the area.

We thank Zeycan & Mustafa of the Terrace Hotel for their co-operation in hosting our event, and the Municipality of Assos and its mayor Hüseyin Kaplan for offering us the guided visit to the ancient site.

As we said, this is the first meeting of its kind, and we certainly hope that it will not be the last!

With best wishes for a successful and happy time here in Assos, from the organizers:

Patrick Groenen
(Erasmus University, Rotterdam, The Netherlands)

Michael Greenacre
(Pompeu Fabra University, Barcelona, Spain)

Zerrin Aşan
(Anadolu University, Eskişehir, Turkey)

Jörg Blasius
(Bonn University, Germany)

Restaurant Les Carmes, Rennes, February 2011

Institutional quality control practices

Jörg Blasius (University of Bonn, Germany),
Victor Thiessen (Dalhousie University, Halifax, Canada)

This paper focuses on procedural and institutional factors that can compromise data quality. Using screening techniques such as MCA, CatPCA, and PCA we were able to detect data problems which one might expect in small studies but not in well-known international surveys such as the world value survey (WVS) or the European Social Survey (ESS). We use three different data sets to illustrate how we detected anomalous patterns and how we subsequently determined that these unusual patterns were, in all likelihood, artefacts of data collection procedures and the failure of data collection institutions to ensure the authenticity of its survey data. The first example comes from the WVS 2005-2008 where we detect clusters of nation-specific response combinations whose frequency of occurrence defies the odds to such an extent that we conclude that some of them occur because of procedural and/or interviewer deficiencies. Next, on the basis of a more stringent definition of anomalous patterns, we document that some data collection institutions probably obtained their quota of cases through the simple expedient of duplicating cases; i.e., the interviews are faked copies, whose existence was masked by making minor changes to the duplicated cases, such as changing the case identifier. This is followed by an example from one of our own primary data collection projects (Blasius et al., 2008) during which our screening procedures detected anomalous patterns associated with the interviews collected by a few of the interviewers. Based on the nature of the patterns, we suspected that these interviewers faked or partially faked some of their interviews. Our suspicions were confirmed through the simple expedient of calling the respondents and determining they had not been interviewed. Finally, we end with an example from the ESS 2002 of a pattern that we conclude represents inadvertent data entry errors.

References

Blasius, Jörg, Jürgen Friedrichs, and Jennifer Klöckner (2008). *Doppelt benachteiligt? Leben in einem deutsch-türkischen Viertel.* Wiesbaden: VS-Verlag.

Student response patterns and scale usage heterogeneity in PISA 2006 science interest scales

Angelos Markos (Democritus University of Thrace, Greece)

Students' interest and attitudes towards scientific issues across the world are generally regarded as central themes in science education and educational policy. The attitudinal component of science is regarded as a psychological trait supporting and maintaining the learning process, and it is highlighted as an important predictor for the choice of a science-related career. The issue of students' interest and attitudes towards science has caught the attention of large-scale, international science assessment projects, such as the Programme for International Student Assessment (PISA), conducted by the Organization for Economic Cooperation and Development (OECD). The PISA 2006 survey of science incorporated conventional Likert-type scales of students' interest towards science.

Analysis of PISA data has indicated a negative correlation between student science interest and science performance at the country level. However, within nearly all countries, students' interest in science was positively related to performance. This apparently paradoxical result may well be an artifact of cross-national differences in response styles. Secondary analysts of these attitudinal data are at risk of reaching erroneous conclusions if they ignore the issue of cultural differences in response style or scale usage heterogeneity, leading to descriptive statistics and inferences that may be biased and misleading. A key question is then to what extent rating scale responses reflect response style and substantive content of the items across PISA nations.

In this talk, I will demonstrate evidence of cross-national response style variation in PISA 2006 science interest scales, using both a set of data screening techniques based on correspondence analysis and categorical principal component analysis (see, for example, Blasius & Thiessen 2006) and a model-based approach developed by Van Rosmalen et al. (2010). I will also illustrate the systematic relationship between the structure of students' reported attitudes and science achievement in each country. The ultimate goal is to contribute to understanding the sources behind the distinct profiles of scientific literacy across the world as expressed across the attitudinal items in PISA.

References

Blasius, J. & Thiessen, V. (2006). A three-step approach to assessing the behaviour of survey items in cross-national research. In: M. Greenacre & J. Blasius (Eds.), *Multiple Correspondence Analysis and Related Methods* (pp. 433–453). Boca Raton, FL: Chapman & Hall.

Van Rosmalen, J.M., Van Herk, H. & Groenen, P.J.F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, **47**, 157 – 172.

Constrained dual scaling of successive categories for detecting response styles

Pieter C. Schoonees, Michel van de Velden, Patrick J.F. Groenen
(Erasmus University Rotterdam, The Netherlands)

Response styles or response bias can arise in questionnaire research when respondents tend to use rating scales in a manner unrelated to the actual content of the survey question (Van Rosmalen, Van Herk, Groenen, 2010). For example, a respondent exhibiting an extreme response style may tend to use the extreme lower or higher ends of the scale regardless of his true preference concerning the item. Such biased responding has been known to occur in cross-national research, for example, and can have a significant impact on the results of a statistical data analysis. Although the possibly detrimental effects of response styles are widely acknowledged, there is no general consensus regarding methods for detecting and purging response styles.

Dual scaling for successive categories (Nishisato, 1980) is a technique related to correspondence analysis (CA) for analyzing categorical data. However, there are important differences between the two techniques in the case of preference data where a respondent has to rate a group of objects simultaneously. One important aspect of dual scaling for successive categories is that it also provides optimal scores for the rating scale. This property as well as the observation that response styles can be interpreted as possibly nonlinear mappings of a group of respondents' latent preferences to a rating scale (Van de Velden, 2007) is used here. Consequently, it is shown that interpreting four different well-known response styles as transformations with specific curvature properties makes it possible to use dual scaling to detect these response styles.

Building on this observation, the relationship between dual scaling and CA in conjunction with non-negative least squares is used to restrict the detected non-linear mappings to conform to monotone spline transformations of the second degree. The impact of these restrictions on the optimal scores is studied and used to identify different response styles by imposing sets of constraints conforming to the different response styles. Simulation studies are utilized to illustrate the effectiveness of the technique and to obtain further refinements. Issues regarding practical applications and significance testing are discussed which paves the way for future developments.

References

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.

Van Rosmalen, J.M., Van Herk, H. & Groenen, P.J.F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, **47**, 157 – 172.

Attribute based and similarity based perceptual mapping methods: Applications on automobile brands and patent data

Ulaş Akküçük (Boğaziçi University, Istanbul, Turkey)

Perceptual mapping is a useful technique that could be employed by marketing managers and researchers in the field of product and brand management. With this technique researchers try to gain insight about the relative positioning of brands with respect to one another. There are a number of multivariate methods that could be used to achieve this purpose, however, they could be summarized under the two broad categories of attribute based and similarity based methods. Each of these categories has different input-output considerations and algorithmic properties. In this paper, the main motivation is to illustrate the differences and similarities between the different perceptual mapping methods with special emphasis on the use of correspondence analysis and multiple correspondence analysis. For the purpose of demonstrating the different algorithms, data have been collected on similarities between automobile brands and attribute ratings on the same brands. In addition, OECD data about patent applications of different countries in various fields have been downloaded. Comparisons are performed using widely available statistical software.

Keywords: Correspondence analysis; perceptual mapping; multidimensional scaling; PROXSCAL; ALSCAL; automobile brands.

Testing and visualizing strategic consensus within and between teams

Murat Tarakci, Nufer Yasin Ates, Jeanine P. Porck, Daan van Knippenberg,
Patrick J.F. Groenen, Marco de Haas (Erasmus University Rotterdam, The Netherlands),

This paper illustrates an application of multivariate data reduction techniques on strategic consensus, the shared understanding of organizational strategy within an organizational unit. Strategic consensus has become a prominent concept in strategy process and strategy implementation research. Yet, research in strategic consensus mostly focuses on the degree of consensus about organizational strategy within a team and does not include other important elements of strategic consensus such as more fine-grained analysis of what different group members agree and disagree on, between-group consensus, or significance testing of differences in consensus (e.g., to evaluate a strategic intervention).

We propose a new analytical approach to study strategic consensus to address these issues and to visualize strategic consensus in an intuitive and easy-to-grasp fashion. First, using vector model of unfolding of the data matrix which has respondents in the columns (as variables) and strategy items (i.e., strategic goals) in the rows (as cases), we represent strategic consensus within a group on biplots where individuals and strategic goals are jointly represented, and so is the strategic consensus. Second, we quantify degree of consensus within and between groups by using object scores of strategic goals obtained in the first step. Third, consensus between groups is visualized by classical multidimensional scaling. Last but not least, we also provide a test of the effectiveness of a consensus-creating intervention via permutation tests.

We demonstrate this methodology in practice using empirical data from a large Western European service provider company. The methodology is closely aligned with the conceptual analyses of strategic consensus and will help research break new ground in more comprehensive analysis of strategic consensus' multifaceted nature. Hence, we conclude with guidelines for research and practice on utilizing the proposed methodology.

Measures of association in comparing multivariate discrete images

Carles M. Cuadras (University of Barcelona, Spain)

In image processing, by varying the wavelength, any material reflects and absorbs the solar radiation in a different way. This is registered by hyperspectral sensors, which collect multivariate discrete images in a series of contiguous wavelength bands, providing the spectral curves, which can distinguish between materials.

In order to partition a multivariate image in regions belonging to different materials, we need to compare these regions which are previously modelled by using compositional data matrices, where the entries in each row is a statistical discrete distribution of the radiance values (columns). These rows correspond to distinct but contiguous wavelengths. Thus the distribution in a row is very similar to the distribution in close rows. To measure this proximity, we use Hellinger distance between rows, which provides a distance matrix.

Given two hyperspectral regions of an image providing two compositional data matrices, we obtain the corresponding distance matrices and, by using metric multidimensional scaling, we compute two sets of principal coordinates, which are related by a multivariate association measure based on canonical correlations.

We illustrate this approach comparing some multivariate regions of images captured by hyperspectral remote sensors.

Keywords: Hyperspectral images; Hellinger distance; multidmensional scaling; canonical correlations; multivariate association.

Some new biplots

Patrick J.F. Groenen (Erasmus University Rotterdam, The Netherlands)

Biplots provide a fantastic tool for visualizing the relations between two entities often with principal component analysis or (multiple) correspondence analysis. Here we discuss three new developments in this context. The first one is the use of the so-called *area biplot*, that can be used as an alternative to every standard projection biplot. Its main difference is that the estimate of the data is given by the area formed by the origin and two points (Gower, Groenen, & Van de Velden, 2010). The second variety is the *nonlinear biplot* with a distance interpretation: the reconstructed value on a variable of each sample point is obtained by finding the nearest marker point on a nonlinear curve representing the variable. This is ongoing joint work with Niel le Roux and Sugnet Lubbe Gardner. The third type of biplot stems from an application of tooth emergence data for school children. The difficulty here lies in the fact that the tooth emergence is not being directly observed, but only intervals are available in which the emergence must have taken place. This *interval-censored* biplot handles this case by providing a biplot and estimating the exact emergence times simultaneously. Each type of biplot will be explained briefly and an example will be presented.

References

Cecere, S., Leroy, R. Groenen, P.J.F., Lesaffre, E., Declerck, D.(in press). Estimating emergence sequences of permanent teeth in Flemish school children using interval-censored biplots. *Community Dentistry and Oral Epidemiology*.

Gower, J.C. and Hand, D.J. (1996). *Biplots*. Monographs on Statistics and Applied Probability, 54. London, U.K.: Chapman & Hall.

Gower, J.C., Groenen, P.J.F. & Van de Velden, M. (2010). Area biplots. *Journal of Computational and Graphical Statistics*, **19**, 46-61.

Gower, J.C., Gardner Lubbe, S., Le Roux , N. (2011). *Understanding Biplots*. Chichester, Wiley.

Analysis of distance biplots: grouped and ungrouped data

Sugnet Lubbe (University of Cape Town, South Africa),
Niël le Roux (Stellenbosch University, South Africa),
John Gower (Open University, United Kingdom)

Asymmetric biplots are the simultaneous graphical representation of the samples and variables in a data matrix. A principal component analysis (PCA) biplot constructed to optimally represent Pythagorean distances between n samples in a low dimensional space can be considered as the simplest form of a biplot. Gower and Hand (1996) generalize this biplot to include different distances between samples (nonlinear biplots) and dissimilarities for categorical variables (generalized biplots).

For grouped data, canonical variate analysis (CVA) biplots, closely related to MANOVA, can be formulated as a two-step procedure. Firstly, the data is transformed such that the Mahalanobis distances between the K group means in the original space become Pythagorean distances in the transformed space. Secondly, a PCA is performed on the K transformed group means. Similar to the decomposition of the total variance in MANOVA, Gower and Krzanowski (1999) generalize CVA to Analysis of Distance (AoD) by decomposing the sum of squared distances and thus relaxing the constraint of equal within class covariance matrices. An AoD biplot can be constructed as a nonlinear biplot of the $K \times K$ dissimilarity matrix of the class means.

In this paper we will show that AoD can be considered to provide the parent methodology with special cases:

- generalized biplots (group sizes = 1, $n = K$)
- nonlinear biplots (generalized biplots with only continuous variables)
- CVA biplots (nonlinear biplot with Mahalanobis distances between samples)
- PCA biplots (nonlinear biplot with Pythagorean distances between samples)
- MCA biplots (generalized biplot with chi-squared distances between samples)

Furthermore, considering AoD as the parent methodology allows for constructing biplots for all the different combinations of continuous, categorical and mixed variables, different dissimilarities and both grouped and ungrouped data. Biplot visualisations (predictive and interpolative) can be made showing: sample points, group centroids, points and/or regions representing within group variation, linear or nonlinear trajectories for quantitative variables and category level points (CLPs) for categorical variables as well as prediction regions for different categories. These different types of biplots will be discussed and examples will be illustrated with the accompanying R software.

Keywords: Biplots; canonical analysis of distance; canonical variate analysis; category level points; multiple correspondence analysis

References

Gower, JC and Hand, DJ. 1996. *Biplots*. Chapman & Hall: London.

Gower, JC and Krzanowski, WJ. 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Applied Statistics*, **48**, 505-519.

Biplots for categorical variables: Grouped and ungrouped data

Niël le Roux (Stellenbosch University, South Africa),
Sugnet Lubbe (University of Cape Town, South Africa),
John Gower (Open University, United Kingdom)

Canonical analysis of distance (AoD) extends canonical variate analysis (CVA) to cope with a wide class of distance functions. We restrict ourselves to Euclidean embeddable distances that are also additive e.g. Pythagorean distance, Clark's distance, Chi-squared distance and the extended matching coefficient. Canonical AoD deals with continuous variables, categorical variables or mixed variables. However, in this paper the focus is on categorical variables.

A key feature of canonical AoD is the analysis of the grouping structure of the data in a space of dimension not larger than $K - 1$ where K denotes the number of groups. Biplots are constructed to represent the sample points, the group centroids and points and/or regions representing within-group variation. The focus of this paper is on representing the categorical variables in the biplot as category level points (CLPs) and prediction regions for different categories.

Generalised biplots give nonlinear biplot trajectories for continuous variables and CLPs for categorical variables. In particular, we are concerned here with the CLPs, which have interesting properties, analogous to those associated with multiple correspondence analysis (MCA), which is a special case. We review these properties and show how many of them extend to the analysis of grouped data by canonical AoD.

R functions to perform canonical AoD for categorical variables and to construct the associated biplots will be introduced. In particular data arising from several fields of application will be used to illustrate the output of these functions.

Keywords: Biplots; canonical analysis of distance; canonical variate analysis; category level points; multiple correspondence analysis

Biplots of fuzzy coded data

Zerrin Aşan (Anadolu Universitesi, Eskişehir, Turkey),
Michael Greenacre (Universitat Pompeu Fabra, Barcelona, Spain)

A biplot, which is the multivariate generalization of the two-variable scatterplot, can be used to visualize the results of many multivariate techniques, especially those that are based on the singular value decomposition. We consider data sets consisting of continuous-scale measurements, their fuzzy coding and the biplots that visualize them, using a fuzzy version of multiple correspondence analysis. Of special interest is the way quality of fit of the biplot is measured, since it is well known that regular (i.e., crisp) multiple correspondence analysis seriously under-estimates this measure. We show how the results of fuzzy multiple correspondence analysis can be defuzzified to obtain estimated values of the original data. The decomposition of variance along principal axes of the defuzzified values is orthogonal and this permits a measure-of-fit to be calculated in the familiar form of a percentage of explained variance per axis, directly comparable to the corresponding measures used in principal component analysis of the original data. The approach is motivated by its application to a simulated data set, showing how the fuzzy approach can lead to diagnosing nonlinear relationships, and we also apply it to a real set of meteorological data.

References

This work has just been published:

Aşan, Z. and Greenacre, M. (2011) Biplots of fuzzy coded data. *Fuzzy Sets and Systems*, **183**, 57 – 71.

Dynamic modifications of multiple correspondence analysis solutions

Alfonso Iodice D'Enza (University of Cassino, Italy)

In several application fields, from social to behavioural sciences, from environmental sciences to marketing, information is gathered and coded in several categorical attributes. In most cases the aim is to identify pattern of associations among the attribute levels. A well-known exploratory method to describe and visualize this type of data is multiple correspondence analysis (MCA) (for a recent account see Greenacre, 2007), the generalization of correspondence analysis (CA) to more than two categorical variables. The MCA implementation consists of an eigenvalue decomposition (EVD) or the related singular value decomposition (SVD) of properly transformed data. The application of EVD and SVD to large and high-dimensional data is not feasible because of the high computational costs and because the whole data structure being decomposed has to be kept in memory. In the literature there are several proposals aiming to overcome the EVD and SVD-related limitations via the update (or down-date) of existing EVD or SVD solutions according to new data. An example of scalable dimension-reduction technique is the incremental PCA proposed by Zhao et al. (2006).

Large categorical data sets are stratified in different batches when they cannot be analysed in a row, or when they refer to information gathered in different occasions in time or space. In these cases a scalable update of a dimension-reduction solution can be suitable to monitor the evolving relationship structures characterizing attributes.

The aim of the present contribution is to propose an MCA-like procedure that can be modified incrementally as new data batches are processed. In particular, the procedure is obtained by integrating a properly modified MCA with an EVD-based approach (Hall et al., 2002) in order to obtain updates and down-dates of the MCA-like solution. Updates will take into account the new data batches analyzed, down-dates will discard older data batches in order to refresh the solution. The low-dimensional quantification and visualization of categorical attributes via this MCA-like procedure is a promising approach to investigate the association structures and for purposes of fast clustering.

Keywords: Multiple correspondence analysis; singular value decomposition update; eigenvalue decomposition down-date and update

References

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

Hall P., Marshall D. and Martin R. (2002). Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition. *Image and Vision Computing*, **20**, 1009-1016.

IodiceD'Enza A. and Greenacre M.J. (2010). Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In: *Proceedings of SIS09, Statistical Methods for the Analysis of Large Data-sets* (in press).

Zhao H., Chi P. and Kwok J. (2006). A novel incremental principal component analysis and its application for face recognition. *Systems, Man and Cybernetics, Part B: Cybernetics, IEEE Transactions*, **35**, 873-886.

Non-symmetrical correspondence analysis for ranked data

Eric J. Beh (University of Newcastle, Australia)
Biagio Simonetti (University of Sannio, Italy)

Correspondence Analysis has been considered under many variations for analyzing different data structures. Greenacre (1984), Nishisato (1980), Weller & Rommey (1990) and Beh (1999) for example, considered several modifications of correspondence analysis, to analyze ranked data. Beh's (1999) approach involved considering Anderson's (1959) statistic, where for ranked data, the Pearson chi-square statistic can be modified such that

$$A^2 = \frac{t-1}{t}\chi^2$$

where t is the number of the treatments ranked by s consumers. Beh (1997) proposed the method of ordinal correspondence analysis which partitions the Pearson chi-square statistic into three components: location, dispersion and higher order components. When calculating these components orthogonal polynomials are defined and require the specification of a scoring scheme to reflect the ordered nature of the variable(s). For example, when ordered ranks are considered, equally spaced integer values may be used. If a simultaneous analysis of the ranks and treatments is considered the following Anderson statistic can be considered:

$$A^2 = n\sum_{u=1}^{t-1}\sum_{v=1}^{t-1} U_{uv}^2$$

where U_{uv} is function of the orthogonal polynomials based on the row and columns marginal totals. In many circumstances, there exists a non-symmetric association between the variables; in this case an appropriate solution is the non-symmetrical correspondence analysis (NSCA) (Lauro & D'Ambra, 1989), based on the tau index (Goodman & Kruskal, 1954; Light & Margolin, 1971). When the dependent variable is an ordinal one, an ordinal non-symmetric correspondence analysis can be applied (Lombardo *et al*., 2007). In this paper we shall be considering a two-way contingence table, where a ranked variable is considered to be the dependent variable.

References

Anderson, R.L. (1959) Use of contingency tables in the analysis of consumer preference data, *Biometrics*, **15**, 582-590.

Beh, E.J. (1999) Correspondence Analysis of Ranked Data, *Commun. Statist. – Theory Meth*., **28**(7), 1511-1533.

Beh, E.J. (1997) Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials, *Biometrical Journal*, **39**, 589-613.

Goodman, L.A. and Kruskal, W.H. (1954) Measures of association for cross classifications, *J. Amer. Statist. Assoc.* **49**, 732-764.

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

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

Light R., Margolin B. (1971) An analysis of variance for categorical data, *J. Amer. Statist. Assoc.* **66** 534–544.

Lombardo R., Beh E. J., D'Ambra L. (2007) Non-symmetric correspondence analysis for doubly ordered contingency tables. *Computational Statistics & Data Analysis*, **52**,1, 566-578.

Nishisato, S. (1980) *Analysis of Categorical Data: Dual Scaling and its Applications*, Toronto Press, Toronto.

Weller, S. and Romney, A.K. (1990) *Metric Scaling: Correspondence Analysis*, Sage University Series of Quantification Applications in Social Sciences, 07-075, Newbury Park, CA: Sage.

Correspondence analysis, chi-square distance and rare objects

Michael Greenacre (Universitat Pompeu Fabra, Barcelona, Spain)

Correspondence analysis is typically applied to a matrix of nonnegative counts such as cross-tabulations encountered in the social sciences. At the heart of the method are the chi-square distance and the weighting of rows and columns proportional to their respective marginal totals in the table. The method is widely applied in linguistics, ecology and archaeology, where the data tables are usually very large as well as sparse, that is the large majority of the entries in the table are zeros. In addition, there are several words (in linguistics), species (in ecology) or artefacts (in archaeology) that occur very "rarely" in the table. Such rare objects are often considered to be problematic in the sense that they appear to be too dominant in the analysis, and various strategies have been proposed to deal with them: for example, combining rare objects with other objects, or simply omitting them from the analysis.

In this talk I will distinguish between two types of "rareness": objects occurring with very low frequency, and objects occurring in very few samples (or a combination of both). I will show that in most cases correspondence analysis copes with rare objects quite naturally and well, thanks to the chi-square distance and the weighting. It is only in the very specific (and rare!) situation that rareness is a problem: namely, when the objects occur in very few samples and when those samples themselves contain only those rare objects.

The talk is illustrated with examples from linguistics and ecology.

References

This paper is available online:

Greenacre, M. (2011). The contributions of rare objects in correspondence analysis. Barcelona GSE Working Paper #571. URL: http://research.barcelonagse.eu/One_Paper.html?paper=571

Objective Bayesian model selection in probit models

Elías Moreno (University of Granada, Spain)

We describe a new variable selection procedure for categorical responses where the candidate models are all probit regression models. The procedure uses objective intrinsic priors for the model parameters, which do not depend on tuning parameters, and ranks the models for the different subsets of covariates according to their model posterior probabilities. When the number of covariates is moderate or large, the number of potential models can be very large, and for those cases we derive a new stochastic search algorithm that explores the potential sets of models driven by their model posterior probabilities. The algorithm allows the user to control the dimension of the candidate models, and thus can handle situations when the number of covariates exceeds the number of observations. Lastly, we assess, through simulations, the performance of the procedure, and apply the variable selector to a gene expression data set, where the response is whether a patient exhibits pneumonia.

Keywords: Intrinsic priors; linear models, Bayes factors, model selection, probit models, stochastic search.

Multidimensional scaling of distances associated with squared correlations and squared partial correlations

Antoine de Falguerolles (Université de Toulouse, France)

The distance between two quantitative statistical variables is frequently derived from the values of their Pearson product-moment correlation coefficient. A well-known formula does the trick and provides Euclidean distances which are not affine invariant. However, in some practical situations, it is thought that the sign of the correlation coefficients should not impact the value of the distances. It turn out that, while the matrix of the absolute values of the correlation coefficients is not necessarily positive definite, the matrix of the squared correlation coefficients is. This result readily extends to the matrix of partial correlation coefficients. Thus, multidimensional scaling of the distances derived from the latter suggests an exploratory approach to the building of the independence graph of the variables. Some features of this proposal will be illustrated on benchmark examples of quantitative variables. The case where the variables are the indicators of the levels of a set of qualitative variables will also be briefly considered.

Developments in population census coverage error methodology

H. Öztas Ayhan (Middle East Technical University Ankara, Turkey)

Coverage error can be defined as the difference between the target population size and the available frame (list) population size. The difference between these is called the *census undercount*. Population census results are used for the allocation of state funds, selection of local representatives, base for representative probability sample selection.

Evidence of undercount (or overcount) in the census can be found by projecting the population for a census year by applying mortality rates and migration rates to the results of other censuses. The pattern of differences between these projections and the actual census counts can provide good evidence for undercount. Research towards undercount adjustments for the case of United States, and for the case of Turkey will be illustrated in detail.

Most widely accepted method for measuring census undercount is called the "*demographic analysis*". The method combines many indicators and develops estimates of national population for each census year by age, race and sex.

Alternative methods are based on "*coverage survey*" (*post enumeration survey*) results to estimate the total population size. "Dual record system" estimates can also be used to obtain an estimate of the total population size.

Measurement error and bias in the undercount adjustment can be quantified by matching error, fabrication of interview, misreporting of usual residence, geocoding errors, and evaluating unreliable interviews. In order to illustrate these issues, numerical examples of coverage amounts will be presented for several countries.

Keywords: Census coverage error; census undercount; population census; survey methodology; undercount adjustment.

Participants: CARME in ASSOS

(ordered by first name)

Alfonso Iodice d'Enza	University of Cassino	Cassino, Italy	iodicede@gmail.com
Angelos Markos	University of Thrace	Alexandroupolis, Greece	amarkos@gmail.com
Antoine de Falguerolles	Université de Toulouse	Toulouse, France	falguero@cict.fr
Biagio Simonetti	Univ. Degli Studi del Sannio	Benevento, Italy	simonetti@usannio.it
Carles Cuadras	University of Barcelona	Barcelona, Spain	ccuadras@ub.edu
Cenk İçöz	Anadolu University	Eskişehir, Turkey	cicoz@anadolu.edu.tr
Cristina Cusani	Accademia delle Belle Arti	Naples, Italy	cristinacusani@gmail.com
Elias Moreno	University of Granada	Granada, Spain	emoreno@ugr.es
Fikret Er	Anadolu University	Eskişehir, Turkey	fer@anadolu.edu.tr
Halil Eryilmaz	Anadolu University	Eskişehir, Turkey	haeryilmaz@anadolu.edu.tr
Harun Sönmez	Anadolu University	Eskişehir, Turkey	hsonmez@anadolu.edu.tr
Jörg Blasius	University of Bonn	Bonn, Germany	jblasius@uni-bonn.de
Levent Terlemez	Anadolu University	Eskişehir, Turkey	lterlemez@anadolu.edu.tr
Magda le Roux		Stellenbosch, South Africa	
Michael Greenacre	Universitat Pompeu Fabra	Barcelona, Spain	michael.greenacre@upf.edu
Murat Tarakci	Erasmus University	Rotterdam, Holland	tarakci@ese.eur.nl
Mª Ángeles (Nines) Murillo		Grenada, Spain	
Niël le Roux	University of Stellenbosch	Stellenbosch, South Africa	njlr@sun.ac.za
Özgür Peker	Anadolu University	Eskişehir, Turkey	opeker@anadolu.edu.tr
Öztaş Ayhan	Middle East Technical University	Ankara, Turkey	oayhan@metu.edu.tr
Patrick Groenen	Erasmus University	Rotterdam, Holland	groenen@ese.eur.nl
Pieter Schoonees	Erasmus University	Rotterdam, Holland	schoonees@ese.eur.nl
Sugnet Lubbe	University of Cape Town	Cape Town, South Africa	Sugnet.Lubbe@uct.ac.za
Ulas Akküçük	Boğazici University	Istanbul, Turkey	ulas.akkucuk@boun.edu.tr
Zerrin Aşan	Anadolu University	Eskişehir, Turkey	zasan@anadolu.edu.tr

IWCA

CARME in ASSOS

Organizing institutions:

ANADOLU UNIVERSITY

UNIVERSITAT POMPEU FABRA

universitätbonn
Rheinische Friedrich-Wilhelms-Universität Bonn

ERASMUS UNIVERSITEIT ROTTERDAM