Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891910776

Contents

Specialized sessions

Solicited Sessions

Contributed papers and Posters

Preface

The COVID-19 pandemic is putting our society under incredible health, emotional, and economic stress. Facing its harmful effects and their uncertainty, the Executive Board of the Italian Statistical Society (SIS) and the Local Organizing Committee, to ensure the highest level of safety for members and delegates, deliberated to cancel the 50th Meeting of the Italian Statistical Society originally planned to be held in Pisa in June 2020 and to postpone the conference to June 2021. The Executive Board and the Local Organizing Committee continue to monitor closely the pandemic evolving situation, and keep the members of SIS and the researchers informed about the potential new dates for the next meeting. To give value to the work of those who prepared their presentation for the conference, the Program Committee decided to publish the volume *Book of short papers - SIS 2020* despite the conference cancellation.

The conference program included 4 plenary sessions, 16 specialized sessions, 24 solicited sessions, 32 contributed sessions and the poster exhibition. Plenary sessions concerned with robust statistics, human longevity, statistical models for climate changes and small area estimation for educational poverty. The meeting had to host also 2 round tables on data privacy and innovation in statistics. Activities focused on topics of interest for a wider audience included two round tables on Teaching Statistics and on the SIS journal Statistical Methods & Applications, and the Stats Under the Stars (SUS6) competition for young statisticians. The SUS6 event attracted many sponsors from statistical, financial and editorial firms as well as numerous students. The conference committee had registered 345 accepted submissions, including 143 to be presented in invited plenary, specialized and solicited sessions, and 202 spontaneously submitted for oral and poster sessions.

This book includes most of the scientific contributions that had to be presented at the 50th Meeting of the Italian Statistical Society. It is organized into 49 chapters corresponding to 15 specialized, 23 solicited sessions, and to 11 general topics for contributed papers and posters. All 268 contributions provide a wide overview of the state-of-the-art of the subjects, from methodological and theoretical contributions, to applied works and case studies. The result is a very lively picture of the Italian statistical community with its international connections.

We would like to thank all contributors for having submitted their work to the conference, the members of the Program Committee and the extra reviewers for their efforts in this difficult period. Although the Conference did not take place, the organization went on until cancellation was decided for safety reasons. It would have been impossible without the joint effort of Università di Pisa, Scuola Superiore Sant'Anna and National Research Council of Pisa. Members these three institutions took part actively in the Local Organizing Committee. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

This book is our contribution to encourage the scientific community and the network of the Italian Statistical Society to go on and transform this difficult period into an opportunity of scientific debate for better statistics in a better world.

Alessio Pollice
Università degli Studi di Bari Aldo Moro
Chair of the Program Committee

Nicola Salvati
Università di Pisa
Chair of the Local Organizing Committee

Francesco Schirripa Spagnolo
Università di Pisa

Program Committee: Alessio Pollice (Chair), Serena Arima, Marilena Barbieri, Alessandra Brazzale, Eugenio Brentari, Alessia Caponera, Antonio Lepore, Antonella Plaia, Tommaso Proietti, MarcoRiani, Nicola Salvati, Pasquale Sarnacchiaro, Mauro Scanu, Manuela Stranges, Valentina Tocchioni, Simone Vantini, Massimo Ventrucci, Paola Vicard, Donatella Vicari.

Local Organizing Committee: Nicola Salvati (Chair), Gaia Bertarelli, Bruno Cheli, Alessandra Coli, Paolo Frumento, Fosca Giannotti, Caterina Giusti, Piero Manfredi, Stefano Marchetti, Lucio Masserini, Vincenzo Mauro, Barbara Pacini, Dino Pedreschi, Francesco Schirripa Spagnolo, Chiara Seghieri.

Organizers of Specialized and Solicited Sessions: Giada Adelfio, Bruno Arpino, Emanuele Aliverti, Nicoletta Balbo, Mara Bernardi, Silvia Bozza, Pierpaolo Brutti. Annalisa Busetta, Michela Cameletti, Carlo Cavicchia, Fabrizio Durante, Leonardo Egidi, Pietro D. Falorsi, Francesco Finazzi, Livio Finos, Stefania Galimberti, Michele Gallo, Caterina Giusti, Francesca Greselin, Alessandra Guglielmi, Francesca Ieva, Tiziana Laureti, Achille Lemmi, Brunero Liseo, Fabio Massimo Lo Verde, Daria Mendola, Roberta Pappadà, Lea Petrella, Alessandra Petrucci, Alessia Pini, Sabrina Prati, Maria Giovanna Ranalli, Davide Risso, Fabrizio Ruggeri, Silvana Salvini, Monica Scannapieco, Francesco Stingo, Luca Tardella, Grazia Vicario, Susanna Zaccarin, Maroussa Zagoraiou.

Non-metric unfolding on augmented data matrix: a copula-based approach

Unfolding non-metrico basato sull'aumento di matrici: un approccio basato sulla funzione copula

Marta Nai Ruscone and Antonio D'Ambrosio

Abstract In this contribution an effective procedure to avoid degeneracies in multidimensional unfolding for preference rank data is proposed. We adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, by using copula-based association measures among rankings (individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). Our proposal is able to both recover the order of the preferences and reproduce the position of both rankings and objects in a geometrical space. Application on real datasets show that our procedure returns non-degenerate unfolding solutions.

Abstract *In questo contributo viene proposta un'efficace procedura per evitare soluzioni degeneri nell'unfolding multidimensionale per dati di tipo rank. La strategia utilizzata é quella di aumentare la matrice dei dati, cercando di costruire una matrice di dissimilaritá completa, utilizzando misure di associazione basate su copule tra rank (individui) e tra rank e oggetti (ovvero, una rappresentazione dell'ordine dei rank e degli oggetti attraverso rank appaiati). La nostra proposta é quindi in grado di catturare sia l'ordine dei rank sia la posizione di rank e oggetti in uno spazio geometrico. Applicazioni su dataset reali mostrano che la nostra procedura restituisce soluzioni non degeneri.*

Key words: copula, unfolding, multidimensional scaling

Marta Nai Ruscone
Universitá di Genova, Via Dodecaneso, 35 - 16146 Genova, e-mail: marta.nairuscone@unige.it

Antonio D'Ambrosio
Universitá di Napoli Federico II, Via Cinthia, M.te S. Angelo - 80125 Napoli, e-mail: antdambr@unina.it

1 The copula function

Copula are functions that join multivariate distribution functions to their marginal distribution functions [8]. They describe the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate normal distribution.

A bivariate copula $C : I^2 \to I$, with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$, is the cumulative bivariate distribution function of a random variable (U_1, U_2) with uniform marginal random variables in [0,1]

$$C(u_1,u_2;\theta) = P(U_1 \leq u_1, U_2 \leq u_2;\theta), \quad 0 \leq u_1 \leq 1 \quad 0 \leq u_2 \leq 1 \tag{1}$$

where θ is a parameter measuring the dependence between U_1 and U_2.

The following theorem by Sklar [8] explains the use of the copula in the characterization of a joint distribution. Let (Y_1, Y_2) be a bivariate random variable with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cdf $F_{Y_1,Y_2}(y_1,y_2;\theta)$, then there always exists a copula function $C(\cdot,\cdot;\theta)$ with $C : I^2 \to I$ such that

$$F_{Y_1,Y_2}(y_1,y_2;\theta) = C\big(F_{Y_1}(y_1), F_{Y_2}(y_2);\theta\big), \quad y_1, y_2 \in \mathbb{R}. \tag{2}$$

Conversely, if $C(\cdot,\cdot;\theta)$ is a copula function and $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are marginal cdfs, then $F_{Y_1,Y_2}(y_1,y_2;\theta)$ is a joint cdf.
If $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous functions then the copula $C(\cdot,\cdot;\theta)$ is unique. Moreover, if $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous the copula can be found by the inverse of (2):

$$C(u_1,u_2) = F_{Y_1,Y_2}(F_{Y_1}^{-1}(u_1), F_{Y_2}^{-1}(u_2)) \tag{3}$$

with $u_1 = F_{Y_1}(y_1)$ and $u_2 = F_{Y_2}(y_2)$. This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The dependence structure is explained by the copula function $C(\cdot,\cdot;\theta)$. Moreover the (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the normal distribution.

Each copula is related to the most important measures of dependency: the Pearson correlation coefficient and the Spearman grade correlation coefficient. The Spearman grade correlation coefficient (see [8] pp. 169-170 for the definition of the grade correlation coefficient for continuous random variables) measures the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula C with parameter θ, then the Spearman grade correlation is

$$\rho_s(C) = 12 \int_{I^2} C_\theta(u_1, u_2) du_1 du_2 - 3. \tag{4}$$

For continuous random variables it is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient [5].

2 Unfolding as a special case of multidimensional scaling on copula-based association between rankings

Unfolding, originally formulated by Coombs [3] for the analysis of the two-mode preference choice data, is a technique that allows the estimation of two configurations usually representing the coordinates for a set of m individuals and a set of n objects on the basis of proximity values between them, typically expressing preferences of each individual over each object.

Therefore unfolding applies multidimensional scaling [4] to an off-diagonal $n \times m$ matrix, usually representing the scores (or the rank) assigned to a set of m items by n individuals or judges [1]. Using of either scores or rankings traditionally discriminates between metric and non-metric unfolding.

The goal is to obtain two configuration of points representing the position of the judges (X) and the items (Y) in a reduced geometrical space. Each point representing the individuals is considered as an ideal point so that its distances to the object points correspond to the preference scores [3].

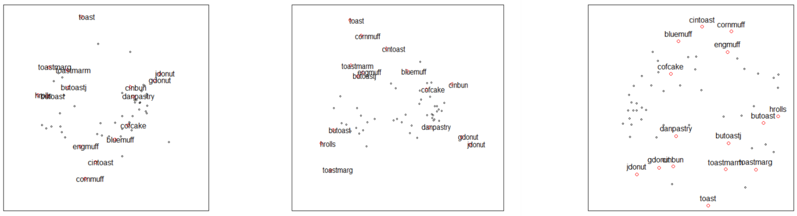
Unfolding can be seen as a special case of multidimensional scaling because the off-diagonal matrix is considered as a block of an ideal distance matrix in which both the within judges and the within items dissimilarities are missing. The presence of blocks of missing data causes the phenomenon of the so-called degenerate solutions, i.e., solutions that return excellent badness of fit measures but not graphically interpretable at all.

To tackle the problem of degenerate solutions, several proposals have been presented in the literature [1]. By following the approach introduced by [9], we adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, and then applying any MDS algorithms.

Let $\boldsymbol{\Gamma}$ be the original $m \times n$ original preference data matrix. In order to augment the data matrix we add to this n additional rows, one for each of the n objects, that correspond to tied rankings representing the jth item, $j = 1, \ldots, n$. As a result, a new $N \times n$ $\boldsymbol{\Gamma}^*$ matrix is obtained, with $N = n + m$. Then we use copula-based association measures among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings), obtaining in fact a $N \times N$ dissimilarity matrix to be analyzed with any MDS algorithm.

3 An application on a real data set

Fig. 1 shows a comparison between the Unfolding solutions of PRESCAL [2], which actually is the most used algorithm for Unfolding analysis, and our proposal by using the Spearman grade correlation coefficient via copula on the breakfast data set. Green and Rao [6] collected 42 rankings of 15 objects by asking 21 students and their wives to order 15 breakfast items in terms of their preference.



(a) Row-conditional Prefscal solution (b) Unconditional Prefscal solution (c) Spearman augmented unfolding

Fig. 1: Unfolding solutions for breakfast data. Breakfast items are labeled as follow: Toast pop-up; Buttered toast; English muffin and margarine; Jelly donut; Cinnamon toast; Blueberry muffin and margarine; Hard rolls and butter; Toast and marmalade; Buttered toast and jelly; Toast and margarine; Cinnamon bun; Danish pastry; Glazed donut; Coffee cake; Corn muffin and butter.

PREFSCAL works by setting two penalties on a modified loss function in such a way to guarantee non degenerate solutions. A possible drawback of this algorithm is that it is not always clear how set the penalty terms. In fact the user must make attempts in order to find the right solution.
The figure emphasizes that the solution of our procedure is not degenerate and it is comparable with the one of PREFSCAL, especially with its unconditional output. It is normal that our output looks like the unconditional PREFSCAL solution because we propose a solution that, depending on how we defined the dissimilarity matrix, is unconditional as well.

4 Concluding remarks

We propose an unfolding algorithm based on the augmentation of the data matrix and a copula-based association between rankings. The shown result highlights that our proposal produces non-degenerate unfolding solutions that are comparable with the ones obtained with PREFSCAL. With respect to PREFSCAL, any parameter must be a priori chosen by the user. On the other hand PREFSCAL always guarantees non degenerate solutions. A robust simulation study will be discussed.

References

1. Borg, I., Groenen, P.: Modern multidimensional scaling. Theory and applications. Springer-Verlag, New York (1997)
2. Busing, Frank M. T. A. and Groenen, Patrick J. K and Heiser, Willem J. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. Psychometrika, **70(1)** 71–98, (2005).
3. Coombs,C.H.: Psychological scaling without a unit measurement. Psychological review, **57** 145–158 (1998)
4. Cox, T.F., Cox, M. A. A.: Multidimensional scaling. Chapman & Hall, London (1994)
5. Embrechts, P., McNeil, A. J., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. Risk management: value at risk and beyond, 176–223 (1998)
6. Green, P E and Rao, V R: Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. Holt. Rinheart and Winston, New York (1972).
7. Kaufman, L., Rousseeuw, P. J.: Finding groups in data. An introduction to cluster analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York (1990)
8. Nelsen, R.B.: An introduction to copulas. Springer Series in Statistics. Springer, New York (2013)
9. Van Deun, K., Heiser W., Delbeke, L.: Multidimensional unfolding by nonmetric multidimensional scaling of spearman distances in the extended permutation polytope. Multivariate Behavioral Research (2007) doi: 10.1080/00273170701341167