Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018
Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors
Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press
fedOA

Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali
Quaderni
11

ASMOD 2018

Proceedings of the International Conference on Advances in Statistical Modelling of Ordinal Data

Naples, 24-26 October 2018

Editors

Stefania Capecchi, Francesca Di Iorio, Rosaria Simone

Federico II University Press

fedOA Press

ASMOD 2018 : Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference : Naples, 24-26 October 2018 / editors Stefania Capecchi, Francesca Di Iorio, Rosaria Simone. – Napoli : FedOAPress, 2018. – (Scuola di Scienze Umane e Sociali. Quaderni ; 11).

Accesso alla versione elettronica:
http://www.fedoabooks.unina.it

ISBN: 978-88-6887-042-3
DOI: 10.6093/978-88-6887-042-3
ISSN Collana: 2499-4774

Università degli Studi di Napoli Federico II
Centro di Ateneo per le Biblioteche “Roberto Pettorino”
Piazza Bellini 59-60
80138 Napoli, Italy
http://www.fedoapress.unina.it/

Published in Italy

Contents

Dissimilarity measure for ranking data via mixture of copulae

Andrea Bonanomi*, Marta Nai Ruscone**, Silvia Angela Osmetti***

Abstract: We propose a new dissimilarity measure for ranking data by using a mixture of copula functions. This measure evaluates the dissimilarity between subjects expressing their preferences by rankings in order to classify them by a hierarchical cluster analysis. The proposed measure is based on the Spearman's grade correlation coefficient on a transformation, operated by the copula, of the rank denoting the level of the importance assigned by subjects in the classification process. The mixtures of copulae are a flexible way to model different types of dependence structures in the data and to consider different situations in the classification process. The advantage by using mixtures of copulae with lower and upper tail dependence is that we can emphasize the agreement on extreme ranks, when extreme ranks are considered more important. An example on simulated data illustrates our proposal.

Keywords: Ranking data, Mixture of copulae, Distance measure.

1. Introduction

Cluster analysis of ranking data aims at the identification of groups of subjects with a homogenous, common, preference behavior. Ranking data occur when a number of subjects are asked to rank a list of objects according to their personal preference order. Cluster analysis input is a distance matrix, whose elements measure the distances between rankings of two subjects. The choice of the distance dramatically affects the final result. The issue when dealing with ordinal data lies in computing an appropriate distance matrix. Several distance measures have been proposed for ranking data (Alvo and Yu, 2014). The most important are referred to Kendall's τ, Spearman's ρ and Cayley distances (Critcholw et *al.*, 1991; Mallows, 1957; Spearman, 1904). When the aim is to emphasize top ranks, weighted distances for ranking data should be used (Tarsitano, 2005). In this context, Bonanomi et *al* (2017) propose a

*Università Cattolica del Sacro Cuore, Milano, andrea.bonanomi@unicatt.it
**LIUC Università Cattaneo, mnairuscone@liuc.it
***Università Cattolica del Sacro Cuore, Milano, silvia.osmetti@unicatt.it

distance measure for ranking data based on copula function with (lower) tail dependence for emphasize the agreement on top ranks, when the top ranks are considered more important than the lower ones.

In this work we propose a generalization of the distance using a mixture of copulae. In this way we have a more flexible instrument to model different types of data dependence structures and to consider different situations in the classification process. For example, by using mixture of copulae with lower tail dependence, we emphasize top ranks or by using mixture of copulae with upper tail dependence, we emphasize low rank. A mixture of copulae with both lower and upper tail dependence permits to assign more weight to both extreme ranks.

An example on simulated data illustrates our proposal.

2. *Our proposal of a dissimilarity measure*

Bivariate copula is a function that captures the dependence structure in a bivariate joint bivariate distribution function. Bivariate copula is, in fact, a class of bivariate distributions, whose marginals are uniform on the unit interval. It describes the dependence structure existing across pairwise marginal random variables (rv).

Sklar's theorem (see Nelsen, 2013) shows that every bivariate/multivariate distribution can be written via a copula representation. Let (Y_1, Y_2) be a bivariate rv with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cumulative distribution function (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

$$F_{Y_1,Y_2}(y_1, y_2; \theta) = C\big(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta\big), \quad y_1, y_2 \in \mathbb{R}. \tag{1}$$

If the marginal cdfs are continuous 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 (1):

$$C(u, v) = F_{Y_1,Y_2}(F_{Y_1}^{-1}(u), F_{Y_2}^{-1}(v); \theta) \tag{2}$$

with $u = F_{Y_1}(y_1)$ and $v = F_{Y_2}(y_2)$.

We consider a new family of copulae defining via finite mixtures (Nelsen,

2013). The idea is to create a new very flexible copula by combining two copulae, as follow:

$$C_M(u, v) = \alpha C_1(u, v; \theta_1) + (1 - \alpha) C_2(u, v; \theta_2) \tag{3}$$

where $\alpha \in [0, 1]$ is the weight of the mixture and C_1 and C_2 are two different copulae, with parameters θ_1 and θ_2, respectively. The two components of the mixture could not be from the same copula family.

We propose to use a mixture of copulae to define the distances between subjects in a hierarchical cluster analysis for ranking data.

We consider two subjects, a and b, expressing their preferences on k objects by rankings.

We consider the mixture of copulae C_M in equation (3) to describe the dependence structure of each pair of latent continuous variables (Y_a^*, Y_b^*) underlying the pair $(Y_a, Y_b) = \{i, j, p_{ij}\}$ for $i, j = 1, 2, .., k$. (Y_a, Y_b) is a bivariate ordinal variable where i and j represent the rank denoting the increasing or decreasing level of the importance assigned to the subjects on the k objects and p_{ij} is the joint frequency with values $1/k$, if the pair (i, j) is observed, and 0, otherwise.

Let F_1 and F_2 the cdfs of Y_a^* and Y_b^*, we assume that each pair (Y_a, Y_b) corresponds to the bivariate discrete random variable obtained by a discretisation of the continuous latent variable $(U = F(Y_a^*), V = F(Y_b^*))$ with support on $[0, 1] \times [0, 1]$, and cdf given by C_M.

Let $A_{ij} = [u_{i-1}, u_i] \times [v_{j-1}, v_j]$, $i, j = 1, 2, \ldots, k$, be the rectangles defining the discretisation. Let $p_{11}, \ldots, p_{kk}$ be the joint probabilities of the ordinal variables corresponding to the rectangles $A_{11}, \ldots, A_{kk}$.

Let $V_{C_M}(A_{11}), \ldots, V_{C_M}(A_{kk})$ be the volumes of the rectangles under the copula C_M, then there exists a unique element in the family of the mixture of copulae that satisfies the following relationship:

$$(V_{C_M}(A_{11}), \ldots, V_{C_M}(A_{ij}), \ldots, V_{C_M}(A_{kk})) = (p_{11}, \ldots, p_{ij}, \ldots p_{kk}). \tag{4}$$

Given the mixture of copulae C_M that satisfies the (4), we define the Spearman's grade correlation coefficients for the pair (Y_a, Y_b), with $a \neq b$, that

performs well in measuring the agreement between two rankings:

$$\rho_S(C_M) = 12 \int_{I^2} [\alpha C_1(u, v; \theta_1) + (1 - \alpha) C_2(u, v; \theta_2)] du dv - 3 \quad (5)$$

The Spearman's grade correlation coefficients of the convex combination of copulae corresponds to the convex combination of the individual Spearman's rho of the two copulae. Finally, the distance $d_{a,b}$ between the rankings of the subjects a and b is:

$$d^C_{a,b} = \sqrt{1 - \frac{\rho_s(C_M) + 1}{2}} \quad (6)$$

We calculate the distances in (6) for each pair of n subjects. We propose to use the obtained $n \times n$ matrix as the dissimilarity matrix in a hierarchical cluster analysis.

By using (6) and the mixture of copulae in hierarchical cluster analysis, we can analyze different situations in the classification process.

For example, we consider three families of Archimedean copulae with different characteristics: Gaussian, Clayton, and Gumbel copula. The Gaussian copula is a symmetric copula that permits positive and negative correlation between the variables and does not allow the dependence in the tails. Instead, Clayton and Gumbel copulae are asymmetric. They permit only positive association and exhibit, respectively, strong left (lower) and right (upper) tail dependence.

By choosing only Gaussian copulae in the mixture, we assign the same weight to all ranks. It is possible to proof that, in a hierarchical cluster analysis, the use of Gaussian copula or classical Spearman rank correlation coefficient (Spearman approach) gives the same classification. By choosing a mixture of Gaussian and Clayton or Gumbel copulae, we can assign to the ranks different "weights" and emphasize the agreement in particular on the top or lower ranks. Therefore, by choosing a mixture of Clayton and Gumbel copulae, we emphasize the agreement only on extreme ranks.

3. An example on simulated data

In this section, we illustrate our proposal by an application to simulated data, analyzed in Bonanomi et *al*. (2017). The data consist on 10 rankings representing the judgements of 10 consumers about 6 aspects of a product, attributing "1" to the most important aspect and "6" to the least important one, reported in Table 1.

Table 1. Example: rankings of 6 products given by 10 consumers

Consumer	*Rankings*						*Consumer*	*Rankings*					
1	1	2	3	4	5	6	**6**	1	2	3	6	5	4
2	2	1	3	4	5	6	**7**	1	2	3	6	4	5
3	1	2	3	4	6	5	**8**	1	2	4	3	5	6
4	2	1	3	4	6	5	**9**	3	1	2	4	5	6
5	3	2	1	4	5	6	**10**	1	2	3	5	4	6

Our aim is to emphasize the extreme ranks (top ranks, lower ranks, or both simultaneously but with different emphasis as well), to develop a more flexible classification than the classical one obtained by Spearman rank correlation coefficient. To achieve this aim, we implement a hierarchical cluster analysis with a distance measure based on a mixture of Gumbel and Clayton copulae. This mixture allows positive associations between rankings and lower and upper tail dependence.

We performed the cluster analysis by using a complete linkage clustering method. We compare the dendrogram obtained by implementing a hierarchical cluster analysis based on the mixture of copulae with the one obtained by the Spearman rank correlation as similarity measure.

The Spearman approach assigns the same importance (weights) at every rank. The mixture of Gumbel and Clayton copulae with weight $\alpha = 0.5$ (equal weight for every copula) assigns to the ranks different weights emphasizing the agreement only on the extreme ranks.

Referring to Table 1, let consider the consumers **1**, **2**, **3** and **8**. If we address the issue of emphasizing top and lower ranks simultaneously, the preferences of consumers **1** and **8** are more similar than **1** and **2**. Moreover, consumer **1**

and **8** would be both separated by consumer **3**.

In Figure 1 we compare the two dendrograms and we show the change of position of the subjects by using the two different approaches.

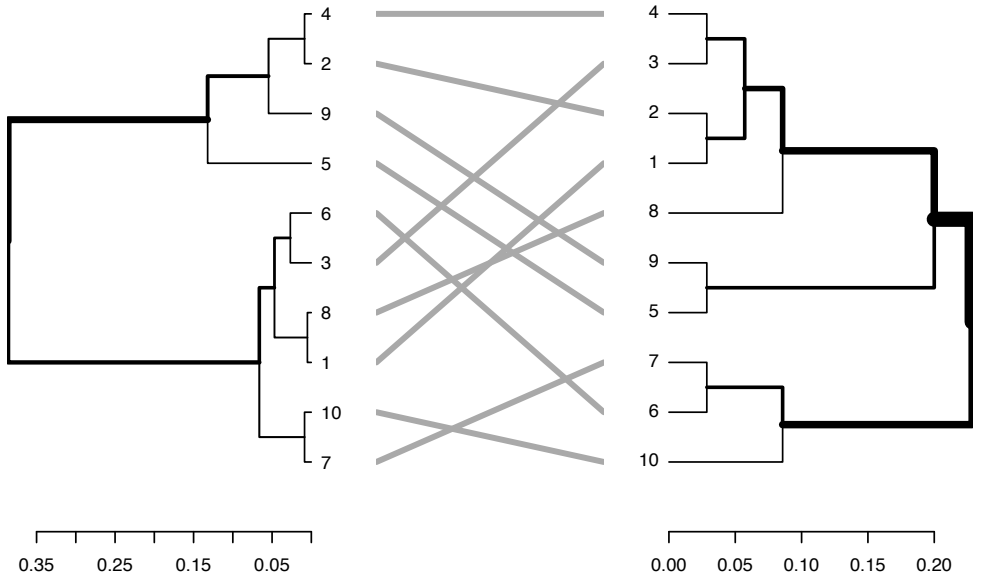

Figure 1. Comparison of dendrograms: Spearman grade correlation coefficient by mixture of copulae (Clayton and Gumbel copulae with weight $\alpha = 0.5$*) (on the left) and Spearman correlation coefficient (on the right)*

The consumers **1** and **8**, whose preferences differ only for the two central ranks, are grouped together at a very low height in the dendrogram obtained by using a mixture (left side of Figure 1), while they are grouped at a greater height in the dendrogram on the right side (Spearman approach).

The consumers **1** and **2**, whose preferences differ only for the two top ranks, are grouped together at a very low height in the dendrogram obtained by using the Spearman approach, while they are grouped at a greater height in the left side.

Moreover, a classification procedure that emphasizes both the top and the lower ranks approaches **1** and **8** and it separates consumer **3**.

In conclusion, the classical approach could be used when one wants to assign equal weights to all ranks in the definition of the distance between rankings. Spearman's grade correlation coefficient ρ_s using the mixture of Gumbel and Clayton copulae gives much more importance on top and lower ranks simultaneously, emphasizing the similarity of consumers with similar extreme ranks.

Acknowledgements: Authors would like to thank the Professor Giuseppe Boari for his useful suggestions.

References

Alvo A., Yu Philip L.H. (2014) *Statistical methods for ranking data*, Springer, New York.

Bonanomi A., Nai Ruscone M., Osmetti S.A. (2017) Defining subjects distance in hierarchical cluster analysis by copula approach, *Quality & Quantity*, 51, 849-872.

Critchlow D.E., Fligner M.A., Verducci, J.S. (1991) Probability models on rankings, *Journal of mathematical psychology*, 35, 294-318.

Mallows C.L. (1957) Non-null ranking models, *Biometrika*, 44, 114-130.

Nelsen R.B. (2013) *An Introduction to Copulas*, Springer, New York.

Spearman C. (1904) The proof and measurement of association between two things, *Am. J. Psychol*, 77-101.

Tarsitano A. (2005) Weighted rank correlation and hierarchical clustering, *Book of Short Paper*, CLADAG2005, Parma.

Università degli Studi di Napoli Federico II
Scuola delle Scienze Umane e Sociali

This volume collects the peer-reviewed contributions presented at the 2nd International Conference on "Advances in Statistical Modelling of Ordinal Data" - ASMOD 2018 - held at the Department of Political Sciences of the University of Naples Federico II (24-26 October 2018). The Conference brought together theoretical and applied statisticians to share the latest studies and developments in the field. In addition to the fundamental topic of latent structure analysis and modelling, the contributions in this volume cover a broad range of topics including measuring dissimilarity, clustering, robustness, CUB models, multivariate models, and permutation tests. The Conference featured six distinguished keynote speakers: Alan Agresti (University of Florida, USA), Brian Francis (Lancaster University, UK), Bettina Gruen (Johannes Kepler University Linz, Austria), Maria Kateri (RWTH Aachen, Germany), Elvezio Ronchetti (University of Geneva, Switzerland), Gerhard Tutz (Ludwig-Maximilians University of Munich, Germany). The volume includes 22 contributions from scholars that were accepted as full papers for inclusion in this edited volume after a blind review process of two anonymous referees.

ISBN: 978-88-6887-042-3
DOI: 10.6093/978-88-6887-042-3
Online ISSN: 2532-4608