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

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

Consensus measures for preference rankings with ties: an approach based on position weighted Kemeny distance

Antonella Plaia*, Mariangela Sciandra **, Simona Buscemi ***

Abstract: Preference data are a particular type of ranking data where some subjects (voters, judges, ...) give their preferences over a set of alternatives (items). It happens, in most of the real cases, that some items receive the same preference by a judge, giving raise to a ranking with ties. The purpose of our paper is to investigate on the consensus between rankings with ties taking into account the importance of swapping elements belonging to the top (or to the bottom) of the ordering (position weights). Combining the structure of the τ_x proposed by Emond and Mason and the class of weighted Kemeny-Snell distances, we propose a position weighted rank correlation coefficient to compare rankings with ties. The one-to-one correspondence between the weighted distance and the rank correlation coefficient holds, analytically speaking, using both equal and decreasing weights.

Keywords: Weighted rank correlation, Weighted Kemeny distance, Position weights.

1. Introduction

Ranking is one of the most simplified cognitive processes useful for people to handle many aspects in their life. When some subjects are asked to indicate their preferences over a set of alternatives (items), ranking data are called preference data. Therefore, ranking data arise when a group of n individuals (judges, experts, voters, raters etc) shows their preferences on a finite set of items (m different alternatives of objects, like movies, activities and so on). If the m items, labeled $1, \ldots m$, are ranked in m distinguishable ranks, a complete ranking or linear ordering is achieved (Cook, 2006): this ranking a is a mapping function from the set of items $\{1, \ldots, m\}$ to the set of ranks

*University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, antonella.plaia@unipa.it

**University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, mariangela.sciandra@unipa.it

***University of Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, simona.buscemi@unipa.it

$\{1, \ldots, m\}$, endowed with the natural ordering of integers, where $a(i)$ is the rank given by the judge to item i. The ranking a is, in this case, one of the $m!$ possible permutations of m elements, containing the preferences given by the judge to the m items. When some items receive the same preference, then a tied ranking or a weak ordering is obtained. In real situations, it can happen that not all items are ranked: partial rankings, when judges are asked to rank only a subset of the whole set of items, and incomplete rankings, when judges can freely choose to rank only some items. In order to get homogeneous groups of subjects having similar preferences, it's natural to measure the spread between rankings through dissimilarity or distance measures among them. Distances between rankings have received a growing consideration in the past few years. Usual examples of metrics in this framework are Kendall's and Spearman's. In 1962 Kemeny introduced a metric defined on linear orders, known as Kemeny distance (or metric), later generalized to the framework of weak orders by Cook et al in 1986, which satisfies the constraints of a distance measure suitable for rankings. The Kemeny distance (d_K) between two rankings a and b is a city-block distance defined as:

$$d_K(a,b) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} |a_{ij} - b_{ij}| \tag{1}$$

where a_{ij} and b_{ij} are the generic elements of the $m \times m$ score matrices associated to a and b respectively, assuming the following values:

$$a_{ij}, b_{ij} = \begin{cases} 1 & \text{if i is preferred to j} \\ 0 & \text{if i = j or if i is tied with j} \\ -1 & \text{if j is preferred to i} \end{cases} \tag{2}$$

Considering the usual relation between a distance d and its corresponding correlation coefficient $\tau = 1 - 2d/D_{max}$, where D_{max} is the maximum distance, d_K is in a one-to-one correspondence to the rank correlation coefficient τ_x

proposed by (Emond and Mason, 2002), defined as:

$$\tau_x(a,b) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} a'_{ij} b'_{ij}}{m(m-1)} \tag{3}$$

where a'_{ij} and b'_{ij} are the generic elements of the $m \times m$ score matrices associated to a and b respectively, assuming the following values

$$a'_{ij}, b'_{ij} = \begin{cases} 1 & \text{if i is preferred to or tied with j} \\ 0 & \text{if i = j} \\ -1 & \text{if j is preferred to i} \end{cases} \tag{4}$$

Distances and correlations are the two possible approach to a consensus ranking problem: given n rankings, full or weak, of m items, what best represents the consensus opinion? This consensus is the ranking that shows the maximum correlation, or equivalently the minimum distance, with the whole set of n rankings.

2. *Weighted distances*

Kumar and Vassilvitskii (2010) introduced two aspects essential for many applications involving distances between rankings: positional weights and element weights. In short, i) the importance given to swapping elements near the head of a ranking could be higher than the same attributed to elements belonging to the tail of the list or ii) swapping elements similar between themselves should be less penalized than swapping elements which aren't similar. In this paper, we deal with case i) and consider the weighted version of the Kemeny metric, since the Kemeny metric is not sensitive towards where the disagreement between two rankings occurs. For measuring the weighted distances, the non-increasing weights vector $w = (w_1, w_2, ..., w_{m-1})$ constrained to $\sum_{i=1}^{m-1} w_i = 1$ is used, where w_i is the weight given to position i in the ranking. Given two generic rankings of m elements, a and b, the weighted Kemeny distance is defined by García-Lapresta and Pérez-Román (2010) as follows:

$$d_K^w(a,b) = \frac{1}{2}\left[\sum_{\substack{i,j=1\\ i<j}}^{m} w_i|a_{ij}^{(\sigma_1)} - b_{ij}^{(\sigma_1)}| + \sum_{\substack{i,j=1\\ i<j}}^{m} w_i|b_{ij}^{(\sigma_2)} - a_{ij}^{(\sigma_2)}|\right], \qquad (5)$$

where (σ_1) states to follow the a ranking and (σ_2), similarly, orders according to b. More specifically, $b_{ij}^{(\sigma_1)}$ is the score matrix of the ranking b reordered according to a, $a_{ij}^{(\sigma_2)}$ is the score matrix of the ranking a reordered according to b and $a_{ij}^{(\sigma_1)} = b_{ij}^{(\sigma_2)}$ is the score matrix of the linear order $1, 2, ..., m$ (see Plaia and Sciandra, 2017 for more details).

3. A new weighted rank correlation coefficient

Recently we proposed a new rank correlation coefficient (Plaia et al, 2018), suitable for position weighted rankings which handles linear orders. In this paper, we propose its generalization to cope with the presence of ties. Combining the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010) and the extension of τ_x provided by Emond and Mason (2002), we propose a new rank correlation coefficient working with a couple of score matrices. Let's define the generic (i, j) element of the score matrices $a_{ij}^{'}$ and a_{ij}^{*} related to a ranking a as follows:

$$a_{ij}^{'}, b_{ij}^{'} = \begin{cases} 1 & \text{if i is preferred to or tied with j} \\ 0 & \text{if i = j} \\ -1 & \text{if j is preferred to i} \end{cases} \qquad a_{ij}^{*}, b_{ij}^{*} = \begin{cases} 1 & \text{if i is preferred to j} \\ 0 & \text{if i = j} \\ -1 & \text{if j is preferred to or tied with i} \end{cases} \qquad (6)$$

Our new rank correlation coefficient uses both these score matrices (the corresponding element of the score matrices are equal to 1 and to -1 according to the considerations in Emond and Mason (2000), secc. 38, 39) and is defined as:

$$\tau_x^w(a,b) = \frac{\sum_{i<j}^{m}(a_{ij}^{'\sigma_1} b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2} b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1} b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2} b_{ij}^{*\sigma_2})w_i}{2Max[d_K^w]}, \qquad (7)$$

where the denominator represents twice the maximum value of the Kemeny weighted distances (García-Lapresta and Pérez-Román, 2010), equal to:

$$Max[d_K^w(a,b)] = 2\sum_{i=1}^{m-1}(m-i)w_i. \tag{8}$$

4. Correspondence between distance and correlation

We will demonstrate that eq. (7) is the correlation coefficient corresponding to the distance (5) through the straightforward linear transformation:

$$\frac{\sum_{i<j}^{m}(a_{ij}^{'\sigma_1}b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2}b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1}b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2}b_{ij}^{*\sigma_2})w_i}{2Max[d_K^w]} = 1 - \frac{2d_K^w}{Max[d_K^w]}$$

or equivalently

$$\sum_{i<j}^{m}(a_{ij}^{'\sigma_1}b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2}b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1}b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2}b_{ij}^{*\sigma_2})w_i = 2Max[d_K^w] - 4d_K^w \tag{9}$$

where $Max[d_K^w]$ and d_K^w are defined in Eq. (8) and in Eq. (5) respectively, and we use the matrix representation of a ranking a of m objects as in Eq. (2) for computing d_K^w and the two different score matrices of Eq. (6) for calculating τ_x^w. According to Emond and Mason (2002), if two rankings a and b agree except for a set S of k objects, which is a segment of both, then $d_K^w(a,b)$ may be computed as if these k objects were the only objects being ranked. As a consequence, to prove the equality in (9) we will show that for each pair of objects i and j:

$$a_{ij}^{'\sigma_1}b_{ij}^{'\sigma_1} + a_{ij}^{'\sigma_2}b_{ij}^{'\sigma_2} + a_{ij}^{*\sigma_1}b_{ij}^{*\sigma_1} + a_{ij}^{*\sigma_2}b_{ij}^{*\sigma_2} = 4(m-i) - 2\left[|a_{ij}^{\sigma_1} - b_{ij}^{\sigma_1}| + |b_{ij}^{\sigma_2} - a_{ij}^{\sigma_2}|\right] \tag{10}$$

In Eq. (10) the weights w_i have been omitted from both the sides. There are nine possible combinations of orderings for item i and j between voters A and B, but only four distinct cases must be considered. The other five are equivalent to one of these four through a simple relabeling of the rankers and/or the objects. (Emond and Mason, 2002).

Case 1. Both A and B prefer object i to j. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_1} = a_{ij}^{\sigma_2} = b_{ij}^{\sigma_2} = 1$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_1} = a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_1} = a_{ij}^{*\sigma_2} = b_{ij}^{*\sigma_2} = 1$. Hence, the equality in

equation (10) holds:
$1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 = 4 - 2[|1 - (1)| + |1 - (1)|]$.
Case 2. A prefers object i to j and B prefers the two objects as tied. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = a_{ij}^{\sigma_2} = 1$ and $b_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = 0$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_1} = a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_1} = a_{ij}^{*\sigma_2} = 1$ and $b_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_2} = -1$. Hence, the equality in equation (10) holds:
$1 \cdot 1 + 1 \cdot 1 + 1 \cdot (-1) + 1 \cdot (-1) = 4 - 2[|1 - 0| + |1 - 0|]$.
Case 3. A prefers object i to j and B prefers j to object i. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = 1$ and $a_{ij}^{\sigma_2} = b_{ij}^{\sigma_1} = -1$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_2} = 1$ and $a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = a_{ij}^{*\sigma_2} = b_{ij}^{*\sigma_1} = -1$. Hence, the equality in equation (10) holds:
$1 \cdot (-1) + (-1) \cdot 1 + 1 \cdot (-1) + (-1) \cdot (1) = 4 - 2[|1 - (-1)| + |1 - (-1)|]$.
Case 4. Both A and B rank the objects i and j as tied. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = a_{ij}^{\sigma_2} = b_{ij}^{\sigma_1} = 0$. The τ_x^w score matrix values are: $a_{ij}^{'\sigma_1} = b_{ij}^{'\sigma_1} = a_{ij}^{'\sigma_2} = b_{ij}^{'\sigma_2} = 1$ and $a_{ij}^{*\sigma_1} = b_{ij}^{*\sigma_1} = a_{ij}^{*\sigma_2} = b_{ij}^{*\sigma_2} = -1$. Hence, the equality in equation (10) holds:
$1 \cdot (1) + (1) \cdot 1 + 1 \cdot (1) + (1) \cdot (1) = 4 - 2[|0 - 0| + |0 - 0|]$.

5. Minimum and Maximum values of τ_x^w

From the demonstrations in sec. 4 τ_x^w can be maximum, and equal to 1, if and only if for all i and j only *Case 1* or only *Case 4* are observed. Therefore, differently from what happens with Kendall τ_b (see Emond and Mason, 2002, sect 3.3), τ_x^w is maximum even when a generic all tied ranking is compared with itself. Analogously, τ_x^w can be minimum, and equal to -1, if and only if for all i and j only *Case 3* occurs.

6. Correspondence between weighted and unweighted measures

For equal weights assigned to the items ($w_i = \frac{1}{m-1}$, for each $i = 1, 2, ..., m-1$) the weighted distance is proportional to the classical Kemeny distance, according to the number of items:

$$d_x^w = \frac{d_x}{m-1}$$

Proof. Referring to the cases listed in Section 4:

Case 1. $d_x^w = \frac{1}{2}[|1-(1)|+|1-(1)|]w_i = 0$ and $d_x = \frac{1}{2}[|0-0|+|0-0|] = 0$

Case 2. $d_x^w = \frac{1}{2}[|1-0|+|1-0|]w_i = \frac{1}{m-1}$ nd $d_x = \frac{1}{2}[|1-0|+|1-0|] = 1$

Case 3. $d_x^w = \frac{1}{2}[|1-(-1)|+|1-(-1)|]w_i = \frac{2}{m-1}$ nd $d_x = \frac{1}{2}[|1-(-1)|+|1-(-1)|] = 2$

Case 4. $d_x^w = \frac{1}{2}[|0-0|+|0-0|]w_i = 0$ and $d_x = \frac{1}{2}[|0-0|+|0-0|] = 0$

Corollary Since $\tau_x \leftrightarrow d_K$ and $\tau_x^w \leftrightarrow d_K^w$, then the weighted rank correlation coefficient is equivalent to the rank correlation coefficient defined by Emond and Mason, when equal importance is given to the positions occupied by the items:

$$\tau_x^w = \tau_x, \quad \text{with } w_i = \frac{1}{m-1} \quad \forall i = 1, 2, ..m-1$$

7. Consensus ranking

The proposed weighted correlation coefficient can be used to deal with a consensus ranking problem: given n rankings, full or weak, of m items, what best represents the consensus opinion? This consensus is the ranking that shows the maximum correlation, with the whole set of n rankings. Given a nxm matrix $\mathbf{X}$, whose l-th row represents the ranking associated to the l-th judge, the consensus ranking, i.e. the ranking c that best represents the matrix $\mathbf{X}$, is that ranking that maximizes the following expression:

$$Max \sum_{l=1}^{n} \frac{\sum_{i<j}^{m}(x_{ij}^{'\sigma_l} c_{ij}^{'\sigma_l} + x_{ij}^{'\sigma_c} c_{ij}^{'\sigma_c} + x_{ij}^{*\sigma_l} c_{ij}^{*\sigma_l} + x_{ij}^{*\sigma_c} c_{ij}^{*\sigma_c}) w_i}{2Max[d_K^w]}$$

8. Conclusions

In this paper, we provided a rank correlation coeffient τ_x^w for weak orderings, as an extension of τ_x^w for linear orderings (Plaia et al, 2018). We demonstrated the correspondence between τ_x^w and the weighted Kemeny distance and, finally, we showed that the weighted rank correlation coefficient τ_x^w is equal to the Emond and Mason rank correlation coefficient τ_x in the case of tied rankings and $w_i = \frac{1}{m-1}$ for all i. Our future purpose is the extension and the implementation in R of the branch and bound algorithm proposed in (Plaia et al 2018) for linear orders to the case of weak orderings.

References

Cook W.D. (2006) Distance based and ad hoc consensus models in ordinal preference ranking, *European Journal Operation Research*, 172, 369-385.

Cook W.D., Kress M., Seiford L.M. (1986) An axiomatic approach to distance on partial orderings, *Operations research*, 20, 115-122.

Emond E.J., Mason D.W. (2000) *A new technique for high level decision support*. Department of National Defence Canada, Operational Research Division, Directorate of Operational Research (Corporate, Air & Maritime).

Emond E.J., Mason D.W. (2002) A new rank correlation coefficient with application to the concensus ranking problem, *Journal of Multi-criteria decision analysis*, 11, 17-28.

García-Laprest J.L., Pérez-Román D. (2010) Consensus measures generated by weighted Kemeny distances on weak orders. In: *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications*, Cairo.

Kemeny J. G. (1962) *Mathematical models in the social sciences*, Ginn and Company.

Kumar R., Vassilvitskii S. (2010) Generalized Distances Between Rankings. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, 571-580, New York, NY, USA. ACM.

Plaia A., Sciandra M. (2017) Weighted distance-based trees for ranking data, *Advances in Data Analysis and Classification*, 1-18. Springer, https://doi.org/10.1007/s11634-017-0306-x.

Plaia A., Buscemi S., Sciandra M. (2018) A proper correlation coefficient for weighted position rankings in a consensus ranking process, *DEASS Working papers*, submitted.

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