ITEM WEIGHTED KEMENY DISTANCE FOR PREFERENCE DATA

Mariangela Sciandra , Simona Buscemi and Antonella Plaia

[1] Department of Economics, Business and Statistics, University of Palermo, (e-mail: mariangela.sciandra@unipa.it, simona.buscemi@unipa.it, antonella.plaia@unipa.it)

ABSTRACT: Preference data represent a particular type of ranking data where a group of people gives their preferences over a set of alternatives. The traditional metrics between rankings don't take into account that the importance of elements can be not uniform. In this paper the item weighted Kemeny distance is introduced and its properties demonstrated.

KEYWORDS: Preference data, item importance, distances.

1 Introduction

Ranking is one of the most simplified cognitive processes used by people to handle many aspects of their lives. 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 for a finite set of items (m different alternatives of objects, like movies, activities and so on). The two representations of a ranking are the rank vector and the order vector. The rank vector lists the ranks given to the objects, the order vector lists the true order of objects in order from best to worst. It is possible to switch from orderings to rankings and vice-versa, and in this paper, we will refer to orderings. If the m items, labeled $1,\dots m$, are ranked in m distinguishable ranks, a complete (full) ranking or linear ordering is achieved (Cook, 2006): this ranking a is a mapping function from the set of items $\{1,\dots,m\}$ to the set of ranks $\{1,\dots,m\}$, endowed with the natural ordering of integers, where $a(i)$ is the rank given by the judge to item i. Ranking a is, in this case, one of the $m!$ possible permutations of m elements. When some items receive the same preference, then a tied ranking or a weak ordering is obtained. In real situations, sometimes not all items are ranked and we talk of 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 obtain homogeneous groups of subjects with similar preferences, it is natural

to measure the spread between rankings through dissimilarity or distance measures d between two rankings, a non-negative value ranging in $0 - Dmax$. In this sense, a consensus is defined as the ranking that is closest (i.e. with the minimum distance) to the whole set of preferences. Another possible way for measuring (dis)-agreement between rankings is in terms of a correlation coefficient: rankings in full agreement are assigned a correlation of $+1$, those in full disagreement are assigned a correlation of -1, and all others lie in between. Kumar and Vassilvitskii (2010) introduced two essential aspects for many applications involving distances between rankings: positional weights and element weights. In brief, i) the importance given to swapping elements near the head of a ranking could be higher than the importance attributed to elements belonging to the tail of the list or ii) changing the ranking of important items should be less penalized than changing the ranking of important ones The first aspect has been widely addressed in literature. Recently Plaia et al (2019a, 2019b) proposed a new position weighted correlation coefficient for linear and weak orderings. Differently, the aspect of element weights is less explored. As Kumar and Vassilvitskii (2010) say, item weights are important, for example, when swapping similar elements should be less penalized than swapping dissimilar elements. To illustrate the idea, when ranking politicians, we should take into account if candidates belong to the same or to different parties: if two rankings differ for the position of two candidates from the same party, it should be reasonable to assume that the distance between these two rankings must be lower than the one between two rankings that differ for the position of candidates that belong to different parties.

In order to take this aspect into account, in this paper we introduce the item weighted Kemeny distance.

2 Distances for ranking data: item weighting

In order to get homogeneous groups of subjects having similar preferences, it is natural to measure the spread between rankings through dissimilarity or distance measures among them. Among the metrics proposed in the literature for computing distances between rankings, we choose to consider the Kemeny distance (Kemeny and Snell, 1962) that, with reference to two rankings a and b, is a city-block distance defined as:

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

a_{ij} and b_{ij} are the generic elements of the $m \times m$ score matrices associated

to a and b, respectively, assuming a value equal of 1 if item i is preferred to item j, -1 if item j is preferred to item i and 0 if the two items are tied or if $i = j$.

The choice of the Kemeny's axiomatic framework (Kemeny and Snell 1962) is justified beacuse we consider the possibility of ties, thus the geometrical space of preference rankings is the generalized permutation polytope (D'Ambrosio and Heiser, 2016), for which the natural distance measure is the Kemeny distance.

In order to consider the possibility that the items are not equally important, we introduce a vector of weights $w = (w_1, w_2, \ldots, w_m)$, with $w_i \geq 0$, whose elements represent the weight (i.e. the importance) we give to each item. The item weighted Kemeny distance is defined as:

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

It is easily demonstrated that the maximum value of eq. (2) is $d_{max} = (m-1)\sum_{i=1}^{m} w_i$.

3 Distance properties

We will prove that eq (2) meet the usual properties of a distance function; given two rankings, a and b:

1. Non negativity: $d(a,b) \geq 0$ and equality hold if and only if $a = b$ limited to items corresponding to weights $w_i > 0$,
2. Symmetry: $d(a,b) = d(b,a)$,
3. Triangle inequality: $d(a,b) \leq d(a,c) + d(c,b)$ if b is between a and c (in case of a metric).

Moreover, a desirable property of any distance is its invariance toward a renumbering of the elements (the so-called label invariance, right invariance or equivariance).

Proof

1. Eq (2) is a sum of absolute values, hence it cannot be negative. If $a \neq b$ at least for the items with corresponding weights greater than 0, then the distance is positive. At the same time, if $a = b$ at least for the items with corresponding weights greater than 0, then the distance is null.
2. Symmetry occurs since $|a_{ij} - b_{ij}| = |b_{ij} - a_{ij}|$.

3. Given i and j, the triangular inequality reduces to:

$$\frac{w_i + w_j}{2}|a_{ij} - b_{ij}| \leq \frac{w_i + w_j}{2}|a_{ij} - c_{ij}| + \frac{w_i + w_j}{2}|c_{ij} - b_{ij}|$$

and dividing by $\frac{w_i+w_j}{2}$ we return to the known Kemeny distance that, as demonstrated by Kemeny and Snell, meets the inequality if c is between a and b.

Finally, since a permutation of items simply rearranges the rows and columns of the score matrix, if a' results from a by a permutation, and b' results from b by the same permutation, then $d_K^{iw}(a',b')$ is the sum of the same terms as $d_K^{iw}(a,b)$, with the terms occurring in a different order: hence the label invariance holds.

References

COOK, W. D. 2006. Distance based and ad hoc consensus models in ordinal preference ranking. *European Journal of Operational Research 369–385*, **172**, 369–385.

D'AMBROSIO, A., HEISER W.J. 2016. A recursive partitioning method for the prediction of preference rankings based upon Kemeny distances. *Psychometrika*, **81**(3), 774–794.

EDMOND, E. J., & MASON, D. W. 2002. A new rank correlation coefficient with application to the concensus ranking problem. *Journal of Multi-criteria decision analysis*.

GARCIA-LAPRESTA, 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., & SNELL, J. L. 1962. *Preference rankings an axiomatic approach*. MIT Press.

KUMAR, R., & VASSILVITSKII, S. 2010. Generalized Distances Between Rankings. *Pages 571–580 of: Proceedings of the 19th International Conference on World Wide Web*. WWW '10. New York, NY, USA: ACM.

PLAIA, A., BUSCEMI, S., & SCIANDRA, M. 2019a. A new position weight correlation coefficient for consensus ranking process without ties. *in press*.

PLAIA, A., BUSCEMI, S., & SCIANDRA, M. 2019b. Consensus among preference rankings: a new weighted correlation coefficient for linear and weak orderings. *submitted*.