On The General Distance Measure

K. Jajuga, M. Walesiak, A. Bak

Wroclaw University of Economics,
Komandorska 118/120,
53-345 Wroclaw, Poland

Abstract: In Walesiak [1993], pp. 44-45 the distance measure was proposed, which can be used for the ordinal data. In the paper the proposal of the general distance measure is given. This measure can be used for data measured in ratio, interval and ordinal scale. The proposal is based on the idea of the generalised correlation coefficient.

1 Introduction

The construction of the particular dependence (e.g. correlation) and distance measure depends on the measurement scale of variables. In the measurement theory four basic scales are distinguished (see e.g. Stevens [1959]): nominal, ordinal, interval and ratio scale. Among them, the nominal scale is considered as the weakest, followed by the ordinal, the interval, and the ratio scale, which is the strongest one. The systematic of scales is based on the transformations that retain the relations of respective scale. These results are well-known and given for example in the paper by Jajuga and Walesiak [2000], p. 106.

2 The Generalised Correlation Coefficient

Consider two variables, say the j-th and the h-th one. A generalised correlation coefficient is given by the following equation (see Kendall and Buckland [1986], p. 266; Kendall [1955], p. 19):

$$\Gamma_{jh} = \frac{\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{ikj}b_{ikh}}{\left[\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{ikj}^2 \sum_{i=2}^{n}\sum_{k=1}^{i-1} b_{ikh}^2\right]^{\frac{1}{2}}}, \tag{1}$$

where: $i,k = 1,...,n$ – the number of objects,
$j,h = 1,...,m$ – the number of variables.

Let us take the vectors of observations $(x_{1j},\ldots,x_{nj})$, $(x_{1h},\ldots,x_{nh})$ on the variables measured on ratio and (or) interval scale. Suppose that a_{ikj}, b_{ikh} are given as:

$$\begin{aligned} a_{ikj} &= (x_{ij} - x_{kj}), \\ b_{ikh} &= (x_{ih} - x_{kh}). \end{aligned} \tag{2}$$

M. Schwaiger et al. (eds.), *Exploratory Data Analysis in Empirical Research*

Then Γ_{jh} becomes Pearson's product-moment correlation coefficient (where x_{ij}, x_{kj} (x_{ih}, x_{kh}) denote i-th, k-th observation on j-th (h-th) variable). The proof is given in Kendall [1955], p. 21.

Let us now take the vectors of observations $(x_{1j}, \ldots, x_{nj}), (x_{1h}, \ldots, x_{nh})$ on the variables measured on ordinal scale. Suppose that a_{ikj}, b_{ikh} are given as:

$$a_{ikj}(b_{ikh}) = \begin{cases} 1 & \text{if} \quad x_{ij} > x_{kj} (x_{ih} > x_{kh}) \\ 0 & \text{if} \quad x_{ij} = x_{kj} (x_{ih} = x_{kh}) \\ -1 & \text{if} \quad x_{ij} < x_{kj} (x_{ih} < x_{kh}) \end{cases} . \tag{3}$$

Then Γ_{jh} becomes Kendall's tau correlation coefficient (Kendall [1955], pp. 19-20). Similarly as Pearson's coefficient, Kendall's tau correlation coefficient takes the values from the interval $[-1; 1]$. The value equal to 1 indicates the perfect consistency between two orders and the value equal to -1 indicates the perfect inconsistency (one order is the inverse of the other one).

In fact, in the Kendall's work in the formula (3) the equality was not considered. We took the more general approach. The value of Kendall's tau coefficient calculated by means of (1) and (3) for raw data is exactly the same as the value of Kendall's tau coefficient calculated by means of the formula (3.3) given in Kendall [1955], p. 35 only for the data for which the ranks were calculated. On the other hand, the application of the formulas (1) and (3) gives the same result for raw data and for the data for which the ranks were calculated. If we use formula by Kendall (formula 3.3 given in Kendall [1955], p. 35) then the observations must be given ranks.

3 The General Distance Measure

Some multivariate statistical methods (for example classification methods, multidimensional scaling methods, ordering methods) are based on the formal notion of the distance between objects (observations). One usually imposes three constraints for the function $d : A \times A \to R$ (A – set of objects, R – set of real numbers) in order to be a distance measure. This function has to be:

- Non-negative: $d_{ik} \geq 0$ for $i, k = 1, \ldots, n$;
- Reflexive: $d_{ik} = 0 \Leftrightarrow i = k$ for $i, k = 1, \ldots, n$;
- Symmetric: $d_{ik} = d_{ki}$ for $i, k = 1, \ldots, n$.

It is easy to notice that the generalised correlation coefficient (including Pearson's and Kendall's coefficient) does not meet the constraints of non-negativity and reflexivity. The constraint of non-negative value can