# 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.

## Keywords

MEASUREMENT SCALES, DISTANCE MEASURES, DATA ANALYSIS

## 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:

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

Then $\Gamma_{jh}$ 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 $\Gamma_{jh}$ 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 be satisfied by using the transformation $d_{ik} = (1 - \Gamma_{ik})/2$ (the values fall into interval $[0; 1]$). However the constraint of reflexivity is still not fulfilled.

We propose here a general distance measure, which meets all three constraints. It is based on the idea of the generalised correlation coefficient. The general distance measure is given by the following equation (see Walesiak [2000]):

$$d_{ik} = \frac{1 - s_{ik}}{2} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{\left[ \sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2 \right]^{\frac{1}{2}}}, \tag{4}$$

where: $d_{ik}(s_{ik})$ – distance (similarity) measure,
$i, k, l = 1, \ldots, n$ – the number of objects,
$j = 1, \ldots, m$ – the number of variables,
$x_{ij}(x_{kj}, x_{lj})$ – $i$-th ($k$-th, $l$-th) observation on the $j$-th variable.

For the variables measured on ratio and (or) interval scale we take $a_{ipj}, b_{krj}$ given as:

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad \text{for } p = k, l \\ b_{krj} &= x_{kj} - x_{rj} \quad \text{for } r = i, l. \end{aligned} \tag{5}$$

Now let us consider the ordinal scale. The only feasible empirical operation on the ordinal scale is counting (the number of the relations: "equal to", "higher than", "lower than"). Therefore in the distance measure we use the relations between the particular object and the other objects.

For the variables measured on ordinal scale we take $a_{ipj}, b_{krj}$ given as (Walesiak [1993], pp. 44-45):

$$a_{ipj}(b_{krj}) = \begin{cases} 1 & \text{if} \quad x_{ij} > x_{pj} (x_{kj} > x_{rj}) \\ 0 & \text{if} \quad x_{ij} = x_{pj} (x_{kj} = x_{rj}) \\ -1 & \text{if} \quad x_{ij} < x_{pj} (x_{kj} < x_{rj}) \end{cases} \quad \text{for } p = k, l; r = i, l \ . \tag{6}$$

Therefore in the denominator of the formula (4) the first factor is the number of the relations "higher than" and "lower than" for object $i$ and the second factor is the number of relations "higher than" and "lower than" for object $k$.

The generalised correlation coefficient is used for the variables, and general distance measure (GDM) for the cases (objects). In the formula for GDM we

used only the idea of the generalised correlation coefficient. The references for the construction of measure (4) with the use of (5) and (6) are respectively Pearson's correlation coefficient (for the variables measured on the interval and ratio scale) and Kendall's tau coefficient (for the variables measured on the ordinal scale). The construction of GDM is based on the relations between two analysed objects and the other objects. This approach is not necessary in the case of the variables measured on the interval and ratio scale, however it is necessary in the case of the variables measured on the ordinal scale. In the case of the ordinal scale the number of the relations: "equal to", "higher than", "lower than" is important, therefore in the construction of the measure the information on the relations between the object and the other objects should be taken into account. The similar method was used in the case of the interval and ratio scale, due to the similarity of the measure (4) to the measure (1).

The measure given as (4) with the use of (5) is applied as the distance measure for the variables measured on the interval and (or) ratio scale. When the formula (6) instead of (5) is used, we get the distance measure for the variables measured on the ordinal scale. Therefore, the distance measure given by (4) cannot be used directly when the variables are measured on different scales. Using (4) and (6) can partially solve this problem, however due to the transformation of data measured on interval and (or) ratio scale into ordinal scale, we loose the information.

## 4 The properties of the general distance measure

The proposed general distance measure $d_{ik}$ has the following properties:

- it can be applied when the variables are measured on the ordinal, interval and ratio scale,
- it takes values from the $[0; 1]$ interval. Value 0 indicates that for the compared objects $i, k$ between corresponding observations of variables, only relations "equal to" take place. If the formula (6) is used, the value 1 indicates that for the compared objects $i, k$ between corresponding observations on ordinal variables, relations "greater than" take place (or relations "greater than" and "equal to") and they are held for other objects (i.e. objects numbered $l = 1, ..., n$ where $l \neq i, k$),
- it satisfies the conditions: $d_{ik} \geq 0, d_{ii} = 0, d_{ik} = d_{ki}$ (for all $i, k = 1, \ldots, n$),
- the empirical analysis proves that distance sometimes does not satisfy the triangle inequality,
- it needs at least one pair of non-identical objects in order to avoid zero in the denominator,
- the transformation of data by any strictly increasing function (formula (6)) or by any linear function (formula (5)) does not change the value of $d_{ik}$.

The distance measure (4) takes care of variables equally weighted. If the weights are not equal then the general distance measure is defined as (see Walesiak [1999]):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj} b_{klj}}{\left[ \sum_{j=1}^{m} \sum_{l=1}^{n} w_j a_{ilj}^2 \sum_{j=1}^{m} \sum_{l=1}^{n} w_j b_{klj}^2 \right]^{\frac{1}{2}}}, \tag{7}$$

and the weights $w_j$ $(j = 1, \ldots, m)$ satisfy conditions $w_j \in (0; m), \sum_{j=1}^{m} w_j = m$.

Three major methods of variable weighting have been developed: an *a priori* method based on the opinions of experts, the procedures based on information included in the data and the combination of these two methods. Gordon [1999], pp. 30-33 and Milligan [1989], pp. 318-325 discuss the problem of variable weighting in multivariate statistical analysis.

We performed simulation study in which the data sets consists of 50 bivariate normal observations representing 4 separated classes. Here the procedures `RNMNGN` and `RNMNPR` were used. They generate the multivariate normal data with given mean vectors and covariance matrices (Brandt [1998], pp. 111-112).

For these data sets the distance matrices were determined by using the distances GDM1 (for the variables measured on the ordinal scale), GDM2 (for the variables measured on the interval scale or the ratio scale), L1 (Manhattan distance), L2 (Euclidean distance) and LN (Chebychev distance). Then the objects were classified by means of four hierarchical methods: average linkage (between groups), average linkage (within groups), nearest neighbour, furthest neighbour. Then it was checked which distances and classification methods lead to the identification of natural clusters. For 12 different data structures and 4 classification methods the best results were obtained in the case when the distances GDM2 and L2 were used.

## 5 Summary

In the paper the general distance measure was proposed. This measure is given by (4) and (5) in the case of the variables measured on the ratio and interval scales and by (4) and (6) in the case of the variables measured on the ordinal scale. The measure is based on the idea of the generalised correlation coefficient. The properties and the results of the simulation studies are also presented. In addition, the computer program `GDM` in the `C++` language, working under Windows 95/98, was written.

**Acknowledgements:** The research presented in the paper was partly supported by the project KBN 5 H02B 030 21.

## References

BRANDT, S. (1998): Analiza danych. Metody statystyczne i obliczeniowe, PWN, Warszawa [Brandt, S. (1997): Statistical and Computational Methods in Data Analysis, Springer-Verlag, New York].

GORDON, A. D. (1999): Classification. Chapman & Hall, London.

JAJUGA, K. and WALESIAK, M. (2000): Standardisation of Data Set Under Different Measurement Scales. In: Decker, R. and Gaul, W. (Eds.): Classification and Information Processing at the Turn of the Millennium. Springer-Verlag, Berlin, Heidelberg, 105-112.

KENDALL, M. G. (1955): Rank Correlation Methods. Griffin, London.

KENDALL, M. G. and BUCKLAND, W. R. (1986): Slownik terminów statystycznych (A Dictionary of Statistical Terms). PWE, Warszawa.

MILLIGAN, G. W. (1989): A Validation Study of a Variable Weighting Algorithm for Cluster Analysis. *Journal of Classification, No. 1, 53-71.*

STEVENS, S. S. (1959): Measurement, Psychophysics and Utility. In: Churchman, C.W. and Ratooch, P. (Eds.): Measurement. Definitions and Theories. Wiley, New York, 18-63.

WALESIAK, M. (1993): Statystyczna analiza wielowymiarowa w badaniach marketingowych [Multivariate Statistical Analysis in Marketing Research]. Wroclaw University of Economics, Research Papers no. 654.

WALESIAK, M. (1999): Distance Measure for Ordinal Data. *Argumenta Oeconomica. No 2 (8), 167-173.*

WALESIAK, M. (2000): Propozycja uogólnionej miary odleglosci w statystycznej analizie wielowymiarowej [The Proposal of the Generalised Distance Measure in Multivariate Statistical Analysis]. In: Paradysz, J. (Ed.): Statystyka regionalna w sluzbie samorzadu lokalnego i biznesu. Wydawnictwo AE, Poznan (in press).