PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1022 2004
TAKSONOMIA 11
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Jajuga, Marek Walesiak
Wrocław University of Economics

REMARKS ON THE DEPENDENCE MEASURES AND THE DISTANCE MEASURES

1. Dependence measure and distance measure

Multivariate statistical analysis methods are very often used in empirical studies. Among the basic types of studies one should mention:

– the studies on the dependence between the variables, where dependence (association) measures are applied,

– the studies on the similarity of multidimensional objects, where distance or similarity measures are applied.

In this paper we present some relations that can be derived for the dependence measures and distance measures. These relations are very well known in the case of classical measures, defined in L_2-norm. We show here that similar relations can be obtained in the case of two other norms, namely: L_1-norm and L_∞-norm.

In the first part as the starting point we take the well-known relation derived between classical correlation coefficient and Euclidean distance. To make the relation meaningful, we consider the distance between two variables rather than between two objects, as in the usual situation met in multivariate statistical analysis. Therefore the observations should be standardized before the calculation of the distance.

Let us denote by n – the number of objects (observations), by m – the number of variables, by r – a correlation coefficient, by d – a distance. The considered relation is valid for the standardized values, where standardization is understood in classical sense, by subtracting the arithmetic mean and dividing by standard deviation (when calculating standard deviation, we divide by the number of observations, rather than by the number of observations minus 1). The considered relation is given by the following formulas (see e.g. [Anderberg 1973]):

$$d_{jk}^{\ 2} = 2n - 2nr_{jk}, \tag{1}$$

$$r_{jk} = 1 - \frac{d_{jk}^{\ 2}}{2n}, \tag{2}$$

where: j, k – the numbers assigned to two variables.

Before looking at the other norms, it is worth to mention that the distance measure, including Euclidean distance, depends on the distribution of random variables. To illustrate this point, let us consider the case of univariate analysis. In this case the distance between two objects (two observations) is simply the absolute value of the difference between two values (the observations on two objects). Since a single variable is considered, there is no need to standardize the values of this variable. Suppose that these two values are equal to –3 and 3, so the distance is equal to 6. However the interpretation whether this distance is large or small, depends on the distribution of the variable. If the distribution is normal with mean 0 and standard deviation 1, we consider this distance as large. If, on the other hand, the distribution is normal with mean 0 and standard deviation 10, we interpret this distance as relatively small. Of course, the similar argument can be put in the multivariate case. So it is fair to make a following statement: The distance measure should be defined with respect to the distribution of the multidimensional vector of variables.

In practice, however, it is often the case, that we do not know the distribution, therefore as a base for the definition of the distance we can take the multidimensional structure of the objects. In this case, the distance between two objects depends not only on the values (given as vectors) for these objects but also on the values of the other objects, to reflect multidimensional structure, that is to reflect the configuration of points in the multidimensional space.

The proposal of such a distance was given by Walesiak, by defining the so-called generalized distance measure (GDM). The derivation and the properties of this measure are given in Walesiak [2002].

Now we adapt this measure to calculate the distance between the variables, of course for standardized values. After some transformations we get the relation between GDM (calculated between variables) and classical correlation coefficient:

$$d_{jk} = \frac{1}{2} - \frac{-4 + m(r_{jk} + 1) - \sum_{\substack{l=1 \\ l \neq j,k}}^{m} r_{jl} - \sum_{\substack{l=1 \\ l \neq j,k}}^{m} r_{kl}}{4 \cdot \left[(m - \sum_{l=1}^{m} r_{jl}) \cdot (m - \sum_{l=1}^{m} r_{kl}) \right]^{0,5}}. \tag{3}$$

From (3) we can conclude the following properties:

– the value of GDM depends on the number of considered variables,

– to calculate this measure, at least for one pair of variables the vectors of values should not be equal (to avoid zero in the denominator),

- when the number of variables is equal to 2, then $d_{12} = 1$,

– the distance between two variables depends on the correlation of these variables with the other variables.

The analysis of the relation (2), which shows how the correlation coefficient between variables is related to the Euclidean distance between these variables in the case of the standardized values, suggests the following algorithm, being the alternative way to calculate classical correlation coefficient:

1. Standardization of values of each variable.
2. Calculation of Euclidean distance between two variables.
3. Transformation to correlation coefficient by using (2) – by virtue of the properties of correlation coefficient it is also the correlation coefficient calculated in the case of non-standardized values.

These considerations lead us to the following problem: What is going to happen if in the presented procedure of the calculation of correlation coefficient, instead of the Euclidean distance, based on L_2-norm, the distance based on the other norm is used? There are at least two other norms, commonly considered in multivariate statistical analysis, namely L_1-norm or L_∞-norm. Then we should possibly get the other dependence measures.

2. Standardization in different norms

To solve the problem described in the previous section, first of all we should define the appropriate standardization. Clearly, the standardization consists in subtracting location parameter and dividing by scale parameter, so that:

$$z_i = \frac{x_i - \mu}{\sigma}. \tag{4}$$

Therefore the standardization in the respective norm should take into account the location parameter and scale parameter appropriate in this norm. To define these parameters, we use some ideas proposed by Jajuga [1999].

First of all, it is worth to notice that the location parameter is the solution to the following problem: minimize with respect to μ the following function:

$$\left(\sum_{i=1}^{n} \left| x_i - \mu \right|^p \right)^{1/p}. \tag{5}$$

Since the above function is based on the L_p–norm, we get the location parameters based on L_p–norm. By assuming different values of p, one gets different possible location parameters. Three of them are well known, namely:

- Median is the solution for p=1;
- Arithmetic mean is the solution for p=2;
- Midrange is the solution for p=∞, where midrange is given as:

$$\mu = 0.5(x_{\max} + x_{\min}). \tag{6}$$

Now we turn to the appropriate definition of scale parameter. Let us note that in the case of L_2–norm:

– location parameter (in this case – arithmetic mean) is the solution to the problem of the minimization of the following function:

$$\sqrt{\sum_{i=1}^{n}(x_i-\mu)^2}\,, \tag{7}$$

– scatter parameter (in this case – standard deviation) is equal to:

$$\sigma=\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(x_i-\mu)^2}\,. \tag{8}$$

So the location parameter is the solution of minimization problem and scatter parameter is the „volume" of the set of observations measured with respect to particular norm (in this case L_2–norm) and using the derived location parameter.

By using the same argument we propose the general form of scatter parameter for L_p–norm:

$$\sigma=\frac{1}{n^{1/p}}\left(\sum_{i=1}^{n}\left|x_i-\mu\right|^p\right)^{1/p}. \tag{9}$$

Therefore, in addition to standard deviation, given by (8), we get the other two scale parameters:

1. For p=1 – the arithmetic mean of the absolute deviations from the median:

$$\sigma=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\mu\right|, \tag{10}$$

2. For p=∞ – the half of range:

$$\sigma=0.5(x_{\max}-x_{\min}). \tag{11}$$

Therefore we get three possible cases of the standardization using (4):

– In L_2–norm – the classical standardization using arithmetic mean and formula (8);

– In L_1–norm – the standardization using median and formula (10);

– In L_∞–norm – the standardization using midrange and formula (11).

It can be proved that for all three cases:

– location parameter (in respective norm) for standardized values is equal to 0;

– scale parameter (in respective norm) for standardized values is equal to 1.

Therefore these properties, well known in the case of L_2–norm, also hold in the other two norms.

3. Other dependence measures

Now we are ready to turn back to the relation between dependence measure and distance measure. After the analysis of Euclidean distance measure it can be proved that:

$$0\le {d_{jk}}^2\le 4n\,, \tag{12}$$

$$0 \le d_{jk} \le 2\sqrt{n}\,. \tag{13}$$

Similarly, after simple calculations we get the related properties for the other two distances:

– In L_1–norm:

$$0 \le d_{jk} \le 2n\,, \tag{14}$$

$$0 \le d_{jk}{}^2 \le 4n^2\,, \tag{15}$$

– In L_∞–norm:

$$0 \le d_{jk} \le 2\,, \tag{16}$$

$$0 \le d_{jk}{}^2 \le 4\,. \tag{17}$$

We can combine formulas (12)-(17) to get the general formula for all three distances:

$$0 \le d_{jk} \le 2n^{1/p}\,, \tag{18}$$

$$0 \le d_{jk}^2 \le 4\left(n^{1/p}\right)^2\,. \tag{19}$$

Using the analogy to the formula (2) and the relation given in (19) we get the formula for the dependence measure defined through distance measure in all three norms. It is given as:

$$r_{jk} = 1 - \frac{d_{jk}{}^2}{2n^{\frac{2}{p}}}\,. \tag{20}$$

Its particular cases are:

– In L_1–norm:

$$r_{jk} = 1 - \frac{d_{jk}{}^2}{2n^2}\,, \tag{21}$$

– In L_∞–norm:

$$r_{jk} = 1 - \frac{d_{jk}{}^2}{2}\,. \tag{22}$$

It can be proved in the case of both coefficients given by (21) and (22):

– these coefficients take values from the interval [–1; 1];
– for exact linear decreasing relationship they take value equal to –1;
– for exact linear increasing relationship they take value equal to 1.

The empirical studies are required to compare the performance of (21) and (22) to classical correlation coefficient.

4. Dependence measure and scale parameter

Now we move to the study of the relation between dependence measures and scale parameter. Here the starting point is the relation between correlation coeffi-

cient and variance (squared standard deviation). Let us denote by *V(X)* variance and by *COV(X,Y)* covariance. It is well known that (see e.g. [Seber 1984]):

$$\frac{V(X_j + X_k) - V(X_j + X_k)}{V(X_j + X_k) + V(X_j + X_k)} = \frac{2COV(X_j, X_k)}{V(X_j) + V(X_k)}. \tag{23}$$

If we standardize each variable by using classical standardization and denote standardized variables by *Z*, we get:

$$\frac{V(Z_j + Z_k) - V(Z_j + Z_k)}{V(Z_j + Z_k) + V(Z_j + Z_k)} = r_{jk}. \tag{24}$$

These considerations lead us to the natural problem. What is going to happen if in formula (24), we replace variance, which is squared scale parameter appropriate for L_2-norm, by other squared scale parameters. There are at least two other norms that should be considered, namely L_1-norm or L_∞-norm. Then we should possibly get the other dependence measures.

We have already derived the appropriate scale parameters for the other two norms. These are:

– For p=1 – the arithmetic mean of the absolute deviations from the median (given in (10)),

– For p=∞ – the half of range (given in (11)).

By applying these scale measures to sums and differences, we get other dependence measures. It can be proved in the case of both coefficients:

- these coefficients take values from the interval [-1; 1];
- for exact linear decreasing relationship they take value equal to –1;
- for exact linear increasing relationship they take value equal to 1.

The empirical studies are required to compare the performance of such dependence measures to the classical correlation coefficient.

References

Anderberg M.R. (1973), *Cluster analysis for applications*, Academic Press, New York.

Jajuga K. (1999), *Some additions to the problem of Lp-norm based parameters*, In: K. Jajuga, M. Walesiak (Ed.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe AE we Wrocławiu nr 817, 23-32.

Seber G.A.F. (1984), *Multivariate observations*, Wiley, New York.

Walesiak M. (2002), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

WARIACJE NA TEMAT MIAR POWIĄZANIA I MIAR ODLEGŁOŚCI

Streszczenie

Artykuł przedstawia pewne propozycje konstruowania miar powiązania, które to miary mogą być traktowane jako konkurencyjne w stosunku do klasycznego współczynnika korelacji. Propozycje te mają u podstaw dwie inne normy. Jedna z propozycji wykorzystuje zależność między kwadratem odległości i współczynnikiem korelacji, zaś druga zależność między parametrem skali a współczynnikiem korelacji.