sciendo

DOI: 10.2478/bile-2019-0015

Biometrical Letters
Vol. 56 (2019), No. 2, 253-261

Comparison of some correlation measures for continuous and categorical data

Ewa Skotarczak, Anita Dobek, Krzysztof Moliński

Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland, ewa.skotarczak@up.poznan.pl

SUMMARY

In the literature there can be found a wide collection of correlation and association coefficients used for different structures of data. Generally, some of the correlation coefficients are conventionally used for continuous data and others for categorical or ordinal observations. The aim of this paper is to verify the performance of various approaches to correlation coefficient estimation for several types of observations. Both simulated and real data were analysed. For continuous variables, Pearson's r^2 and MIC were determined, whereas for categorized data three approaches were compared: Cramér's V, Joe's estimator, and the regression-based estimator. Two method of discretization for continuous data were used. The following conclusions were drawn: the regression-based approach yielded the best results for data with the highest assumed r^2 coefficient, whereas Joe's estimator was the better approximation of true correlation when the assumed r^2 was small; and the MIC estimator detected the maximal level of dependency for data having a quadratic relation. Moreover, the discretization method applied to data with a non-linear dependency can cause loss of dependency information. The calculations were supported by the R packages *arules* and *minerva*.

Keywords: correlation, mutual information, contingency table

1. Introduction

The measure of correlation is usually one of the starting points in a multivariate data analysis. In the literature there can be found a wide collection of correlation and association coefficients used for different structures of data. Kurt et al. (2016) give a comprehensive review of correlation coefficients, including their comparison. They compared the correlation coefficients used in the inference of gene networks. The described coefficients are well known and are also commonly applied in other fields. Generally, some of the correlation coefficients are

conventionally used for continuous data, and others for categorical or ordinal observations. Pearson's r^2 is the basic measure of linear correlation for normally distributed data, whereas the Spearman coefficient is appropriate for ranks. When for some reason the collected observations have been discretized, forming a contingency table, the estimation of relations between the original variables becomes more difficult.

Some association measures based on chi-squared statistics are available for contingency tables, i.e. Cramér's *V*, *phi*, or Cohen's *h* for proportions (Cramér, 1946; Sheskin, 2004). Another approach to the estimation of correlation has developed out of information theory. The maximal information coefficient (MIC) proposed by Reshef et al. (2011) is presented as a very useful tool in the detection of linear and non-linear relationships between continuous variables.

To answer the question how close is the correlation of categorical data to the correlation of continuous data, Joe (1989) described the relation between Pearson's correlation and a relative entropy measure of multivariate dependence. Skotarczak et al. (2018) developed a regression-based approach to the assessment of correlation coefficients using normalized mutual information. Simulated studies conducted for two-row contingency tables demonstrated some advantages of the regression approach over Joe's estimator.

The aim of the present paper is to test the performance of various approaches to correlation coefficient estimation for several types of observations. Both simulated and real data were analysed. For the continuous variables, Pearson's r^2 and MIC were calculated, whereas for categorized data three approaches were compared: Cramér's *V*, Joe's estimator, and the regression-based estimator.

2. Material and methods

The formula for calculating Pearson's r^2 is well known. We give below a short presentation of the other correlation measures used in this paper. Cramér's *V* is a measure based on chi-squared statistics of dependency, calculated according to the formula

$$V = \sqrt{\frac{\chi^2 / N}{\min(r-1, c-1)}},$$

where N is the total number of observations classified in the contingency table, r is the number of rows and c is the number of columns of the table. Cramér's V lies in a range from 0 to 1. It takes the value zero if and only if $\chi^2 = 0$, hence $V = 0$ is interpreted as independence, whereas $V = 1$ indicates perfect association (Cramér, 1946).

To introduce the other correlation measures used in this paper, based on the concepts of entropy and mutual information, let us recall the entropy formula (Shannon, 1948). For a categorical random variable **A** taking values $\{a_1, a_2,\ldots,a_k\}$ with probabilities $p(a_i)$ ($i = 1,2,\ldots,k$), the entropy is defined as

$$H(\mathbf{A}) = -\sum_{i=1}^{k} p(a_i)\log(p(a_i)).$$

For two variables **A** and **B** this formula can be expressed as

$$H(\mathbf{A},\mathbf{B}) = -\sum_{i=1}^{k}\sum_{j=1}^{l} p(a_i, b_j)\log p(a_i, b_j),$$

where a_i ($i = 1,2,\ldots,k$) and b_j ($j = 1,2,\ldots,l$) are the values of **A** and **B**, and p is the corresponding probability (Jakulin, 2005). Furthermore, the conditional entropy describing the uncertainty about variable **A** in the presence of **B** can be calculated according to the following rule:

$$H(\mathbf{A}/\mathbf{B}) = -\sum_{j=1}^{l} p(b_j)\sum_{i=1}^{k} p(a_i / b_j)\log p(a_i / b_j).$$

The mutual information between variables **A** and **B** is then defined as

$$\begin{aligned} I(\mathbf{A},\mathbf{B}) &= H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A},\mathbf{B}) = H(\mathbf{A}) - H(\mathbf{A}/\mathbf{B}) = H(\mathbf{B}) - H(\mathbf{B}/\mathbf{A}) \\ &= H(\mathbf{A},\mathbf{B}) - H(\mathbf{A}/\mathbf{B}) - H(\mathbf{B}/\mathbf{A}). \end{aligned}$$

The mutual information is the basic value used to construct Joe's correlation estimator, the regression-based correlation estimator and the MIC estimator. In Joe's method, Pearson's correlation coefficient is approximated by $\sqrt{J}$, where

$$J(\mathbf{A},\mathbf{B}) = \frac{I(\mathbf{A},\mathbf{B})}{\min\{H(\mathbf{A}), H(\mathbf{B})\}}$$

(Joe, 1989). The regression approach was developed for simulated data by regressing the linear dependency between –ln(r) and –ln(J), where ln denotes the natural logarithm. The best approximation of the assumed correlation was achieved taking $r^2 = \exp(-b_0)J^{b_1}$, where b_1 and b_0 are regression coefficients whose values depend on the size of the contingency table for the analysed data. The values of b_1 and b_0 for two-row tables with two, three, four and five columns are given in a paper by Skotarczak et al. (2018).

The MIC coefficient is another correlation measure based on the mutual information; however, it is defined for continuous data. The estimation of mutual information for continuous variables is based on the idea that for variables **A** and **B** a grid can be drawn on a scatterplot of **A** and **B**, where the values of variables are assigned to the appropriate row and column of the grid (Reshef et al., 2011). The mutual information values are then calculated by exploring all possible grids up to a maximal grid resolution. The definition and properties of MIC were described by Reshef et al. (2011) and discussed by Kinney et al. (2013) and Kurt et al. (2016). MIC takes values from 0 to 1; the closer it is to 1, the stronger relationship between the variables is expected. To calculate MIC we used the functions available in the *minerva* package in R (Albanese et al., 2012).

In the present work the presented correlation measures were applied to simulated as well as real examples. To begin with, two vectors of variables **X** and **Y** each of size n = 300 were generated from a two-dimensional normal distribution with zero expectations and a given value of linear correlation (r^2). Two levels of correlation were used: 0.1 and 0.9. In the next step, the continuous variables were discretized and classified into two-row contingency tables. The discretization was performed with the *discretize* function from the R package *arules* (Hahsler et al., 2011), using two methods: “interval” (equal interval width) or “frequency” (equal frequency). Six sets of simulated data are displayed in Table 1.

Table 1. Simulated data sets

data1, r^2=0.9, discretization method "interval"					data3, r^2=0.1, discretization method "interval"				
class	1	2	3	4	class	1	2	3	4
1	15	129	47	0	1	10	57	84	15
2	0	7	92	10	2	5	39	67	23
data2, r^2=0.9, discretization method "frequency"					data4, r^2=0.1, discretization method "frequency"				
class	1	2	3	4	class	1	2	3	4
1	72	59	18	1	1	39	45	34	32
2	3	16	57	74	2	36	30	41	43
data5, quadratic dependency, discretization method "frequency"					data6, quadratic dependency discretization method "interval"				
class	1	2	3	4	class	1	2	3	4
1	38	37	37	38	1	156	5	3	2
2	37	38	38	37	2	122	6	3	3

In the sets denoted data1 and data2, the assumed correlation between **X** and **Y** was 0.9, while for data3 and data4 this correlation was 0.1. In data1 and data3 the discretization was performed using the equal interval method, and in data2 and data4 it used the equal frequency method. The tables denoted data5 and data6 were collected following discretization of the continuous variables **X** and $\mathbf{X}^2$, where $\mathbf{X}^2$ was calculated simply by squaring each value of **X**. In data5 the equal frequency method was used in the discretization process, and in data6 the equal interval method was used.

Some contingency tables from real experiments available in the literature were also analysed; these are listed in Table 2.

Table 2. Experimental data sets

data7	children	
preferred doll	Afro-american	White
black	62	11
white	27	60

data8	success in second year	
success in first year	yes	no
yes	46	35
no	34	56

data9	year				
nest fate	2000	2001	2002	2003	2004
hatched	20	14	20	22	19
failed	8	19	34	30	16

The table denoted data7 is taken from a paper by Hraba and Grant (1970) investigating racial preference among Nebraska school children in the year 1969, and data8 is the contingency table describing second year nesting success for successful and unsuccessful nests in the first year (Savard, 1988). The last table, denoted data9, is taken from a paper by Savard and Robert (2007), and contains data on the nesting success of goldeneyes in nest boxes installed in fall 1998 and 1999 in the boreal forest of Quebec, Canada.

For the generated continuous variables, the maximal information coefficient MIC was calculated and compared with the known r^2. Moreover, the ability of MIC to detect a quadratic relationship between variables was also investigated; i.e. MIC was calculated between the variables $\mathbf{X}$ and $\mathbf{X}^2$. For observations collected in a contingency table, the chi-squared independence test was performed and the above-mentioned correlation measures were calculated: Cramér's *V*, Joe's entropy-based estimator, and the regression-based estimator.

The calculations were performed using the R package (R Core Team, 2013).

3. Results

The values of estimated correlation measures for the nine data sets are presented in Table 3.

Table 3. Assumed and calculated correlation measures

Data set	r^2_{assumed}	MIC	χ^2 (*p*-value)	Cramér's *V*	JC	RBC
data1	0.9	0.725	136.82 (0.0005)	0.675	0.638	0.868
data2	0.9	0.725	179.47 (<0.0001)	0.773	0.727	0.991
data3	0.1	0.192	5.287 (0.152)	0.133	0.113	0.149
data4	0.1	0.192	5.387 (0.145)	0.134	0.114	0.151
data5	-	1.0	5.822 (0.087)	0.139	0.184	0.246
data6	-	1.0	0.053 (0.997)	0.014	0.011	0.014
data7	-	-	44.556 (<0.0001)	0.528	0.477	0.666
data8	-	-	5.449 (0.020)	0.178	0.161	0.322
data9	-	-	10.342 (0.035)	0.226	0.194	0.231

r^2 – Pearson's correlation coefficient, MIC – maximal information criterion, χ^2 – value of test statistics in chi-squared test of independency, JC – Joe's estimator, RBC – regression-based criterion

For data with the high assumed linear correlation coefficient, i.e. data1 and data2, the estimators closest to the true correlation value were those based on the regression approach. Other calculated measures had smaller values than the assumed r^2. Cramér's *V* and Joe's estimator were slightly closer to the true value of the correlation when discretization was performed with the equal frequency method. Also, MIC underestimated the assumed value of correlation in this case.

For the sets data3 and data4, for which the assumed correlation coefficient was small ($r^2_{assumed} = 0.1$), the calculated measures showed a reverse tendency: the regression-based estimator overestimated the assumed value, whereas Cramér's *V* and Joe's measure were close to the true correlation. It seems that the discretization method did not influence the obtained estimators for weakly correlated data. MIC overestimated the assumed r^2 for these variables.

The sets data5 and data6 contained data with the quadratic relationship discretized using two methods. It is notable that in this case, the calculated MIC value was the maximum 1.0, detecting the strong functional dependency between the variables. However, after discretization, particularly with the use of the equal interval method (data6), all information about the dependency between the generated variables was lost.

For the data from real experiments, denoted data7, data8 and data9, all of the calculated coefficients proved to have the potential to show the dependency between the observations. When the dependency was proved with a small p-value in the chi-squared test, the values of Cramér's *V*, Joe's estimator and the regression-based estimator were highest. They became smaller with higher p-values (results for data7 compared with data8 and data9). MIC was not calculated for these cases because the data were available only as contingency tables.

4. Conclusions

Observations collected into a contingency table can be determined by unobservable variables of a continuous nature. This assumption is made, for example, for threshold traits and especially for binary traits, when only two stages

of the trait – success or failure – are recorded, although each stage is conditioned by an unobserved liability with continuous distribution (Harville and Mee, 1984). For two threshold traits observed on an experimental unit, it would be valuable to know the correlation of the underlying continuous variability. Measures of this correlation have recently been developed. One class of these measures is based on information theory. In this paper we used three information-based measures of correlation: MIC for continuous variables, and Joe's estimator and a regression-based estimator for two-row contingency tables.

Characteristics of Joe's estimator and the regression-based estimator for simulated data sets were investigated in our earlier paper (Skotarczak et al., 2018). Under many generated data scenarios it was proven that the regression approach yields more accurate results than Joe's proposal, especially for highly correlated data. The current analysis confirmed these conclusions. The regression-based approach yielded the best results for data with the highest assumed r^2 coefficient, whereas Joe's estimator was the better approximation of the true correlation when the assumed r^2 was small. The MIC estimator detected the maximal level of dependency for data having a quadratic relation. However, the regression-based measure was also larger than zero (in the case with the equal frequency discretization method) and was higher than Cramér's V and Joe's estimator. The discretization method applied to data with a non-linear dependency can cause loss of dependency information. Compared with classical dependency measures based on chi-squared statistics, such as Cramér's V coefficient, the information-based correlation measures are worthy of consideration as means to estimate correlation between continuous as well as categorized variables.

REFERENCES

Albanese D., Filosi M., Visintainer R., Riccadonna S., Jurman G., Furlanello C. (2012): Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics 707.

Cramér H. (1946): Mathematical Methods of Statistics. Princeton: Princeton University Press.

Harville D.A, Mee R.W. (1984): A mixed model procedure for analyzing ordered categorical data. Biometrics 40: 393-408.

Hraba J., Grant G. (1970): Black is beautiful: a reexamination of racial preference and identification. Journal of Personality and Social Psychology, 16 (3): 398-402.

Hahsler M., Chelluboina S., Hornik K., Buchta C. (2011): The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. Journal of Machine Learning Research 12: 2021-2025.

Jakulin A. (2005): Machine learning based on attribute interactions. PhD dissertation. University of Ljubljana.

Joe H. (1989): Relative entropy measures of multivariate dependence. Journal of the American Statistical Association 84(405): 157-164.

Kinney J.B., Atwal G.S. (2013): Equitability, mutual information, and the maximal information coefficient. arXiv: 1301.7745.

Kurt Z., Aydin N., Altay G. (2016): Comprehensive review of association estimators for the inference of gene networks. Turkish Journal of Electrical Engineering & Computer Sciences 24: 695-718.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reshef D.N., Reshef Y.A., Finucane H.K., Grossman S.R., McVean G., Turnbaugh P.J., Lander E.S., Mitzenmacher M., Sabeti P.C. (2011): Detecting novel associations in large data sets. Science 334(6062): 1518-1524.

Savard J-P.L., Robert M. (2007): Use of nest boxes by goldeneyes in Eastern North America. The Wilson Journal of Ornithology 119 (1): 28-34.

Savard J.L. (1988): Use of nest boxes by Barrow's goldeneyes: nesting success and effect on the breeding population. Wildlife Society Bulletin 16: 125-132.

Shannon C.E. (1948): A mathematical theory of communication. The Bell System Technical Journal (27): 379-423, 623-656.

Sheskin D.J. (2004): Handbook of parametric and nonparametric statistical procedures. CRC Press, Boca Raton.

Skotarczak E., Dobek A., Moliński K. (2018): Entropy as a measure of dependency for categorized data. Biometrical Letters 55(2): 233-243.