Divya PANDOVE, Shivani GOEL, Rinkle RANI, 2018. An intuitive general rank-based correlation coefficient. *Frontiers of Information Technology & Electronic Engineering*, 19(6):699-711. https://doi.org/10.1631/FITEE.1601549

An intuitive general rank-based correlation coefficient

Key words: General rank-based correlation coefficient ; Multivariate analysis; Predictive metric; Spearman's rank correlation coefficient

Corresponding author: Divya Pandove
E-mail: dpandove@gmail.com
ORCID: https://orcid.org/0000-0001-8694-1538

Motivations

1. Increasement in computational capacity and analytical capabilities has led to expansion in the area of data analysis.

2. The data sets are becoming large and more complex.

3. Hypothesis' are no longer driven by trial and error methods.

4. The fuller proof method is to select a proxy to study relationships between two variables and verify how good the proxy is by using correlation analysis.

5. Instead of a 'hypothesis driven approach', a 'data driven approach' is used.

6. This approach forms the basis of predictive analysis as it relies on good proxies to predict future events. The focus of the analysis is not on the cause of the problem but just indicating that the problem exists.

Main ideas

1. To correctly predict correlation values in large data sets, there is a need to refine the existing correlation determining factors to include the various problems encountered in big data.

2. These are the problems of dimensionality, outliers, and coefficients showing fake correlations which do not exist.

3. In this paper, the focus is on subsets of observations which have same multivariate features.

4. There is a need to perform multivariate feature selection and identification of predictive set of metrics which best suit our purpose.

5. In this work a predictive metric is proposed, defined by a series of steps, to calculate correlations between paired values.

6. This metric is called the ‘general rank based correlation coefficient’ (GRCC).

Methods

1. A predictive metric GRCC is proposed. This metric finds correlations among two variables *x* and *y*.

2. The metric *g* is governed by a parameter *c* which is a prior distribution. This metric is considered only at *c*=1 and *c*=2 and is symmetric in nature.

3. This means that it will not change even if *x* and *y* are swapped. If the order of *y* is reversed then only the sign of correlation will change. Also for every value of *c*>0 the value of *g* will always lie between −1 and 1.

4. When value of *c* is equal to 2, it becomes very sensitive to outliers. The value of *g* at *c*=1 is the most well rounded solution. Further, rank distance between *x* and *y* is calculated in *a* and between *x* and re-verse order of *y* in *b*. Mathematically, *g* can be expressed as:

$$g = s[1 - \min(a, b)/d]$$

Methods

1. The smallest value between a and b helps to determine the sign(s) of the correlation. The value of the denominator d is the most crucial to determine the exact value of g.

2. There are 3 ways mentioned in the technique, any one of them can be used to determine the value of d depending on various factors such as the number of data points.

3. The value of the denominator d can be selected in any of the ways given in Fig. 4.

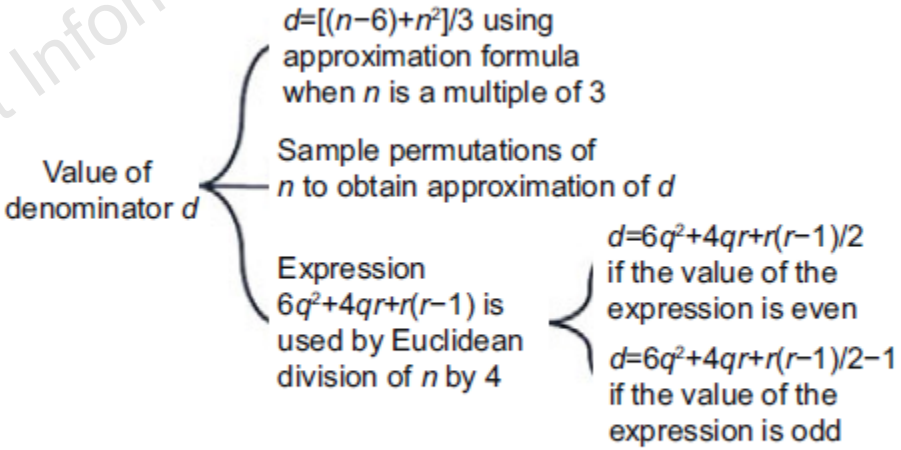

Fig. 4 Ways to select the value of denominator d

Major results

1. To test the proposed approach, a world bank data set is selected. This data set is tested on a data model created by g. The results have been compared with the Spearman's rank correlation. The comparative analysis is given in table 1.

Table 2 Comparison of Spearman's rank correlation coefficient and GRCC

Data attribute		Number of data points	Spearman's rank correlation coefficient			GRCC g			
						$d = d_1$		$d = d_2$	
A	B		p value	S value	ρ	$c = 1$	$c = 2$	$c = 1$	$c = 2$
GDP per capita	GDP	2327	$< 2.2 \times 10^{-16}$	900 579 181	0.4600	0.4302	0.3216	0.4130	0.3092
GDP per capita	Passenger cars	2354	$< 2.2 \times 10^{-16}$	1 451 827 312	0.3321	1.0000	1.0000	1.0000	1.0000
Life expectancy	Health expenditure per capita	2327	$< 2.2 \times 10^{-16}$	858 399 753	0.5380	0.9999	0.9984	0.9997	0.9842
Passenger cars	Urban population	2354	$< 2.2 \times 10^{-16}$	1 365 948 542	0.3717	0.7480	0.9890	0.7030	0.9420
GDP per capita	Internet users	2354	$< 2.2 \times 10^{-16}$	1 101 037 602	0.4935	0.7551	0.8821	0.8155	0.8122
Mobile subscribers	Internet users	2354	$< 2.2 \times 10^{-16}$	224 906 329	0.8965	0.7551	0.6128	0.4130	0.8551
Under-5 mortality	Birth rate	2354	$< 2.2 \times 10^{-16}$	120 426 906	0.9440	0.5110	0.4626	0.4780	0.4400
Total population	Urban population	2354	$< 2.2 \times 10^{-16}$	104 317 436	0.9520	0.4303	0.3416	0.3193	0.3253

$d_1 = 6q^2 + 4qr + r(r-1)/2, d_2 = [(n-6) + n^2]/3$

2. g lies between -1 and 1, hence, it is bounded. It has a bimodal distribution with a small dip near 0. This means that near 0, no patterns would be found. It can be used for detection of outliers as it uses rank distances instead of squared distances.

Major results (Cont'd)

The proposed metric has higher correlations among the variables which should have a high correlation according to the human understanding of the world and vice versa.

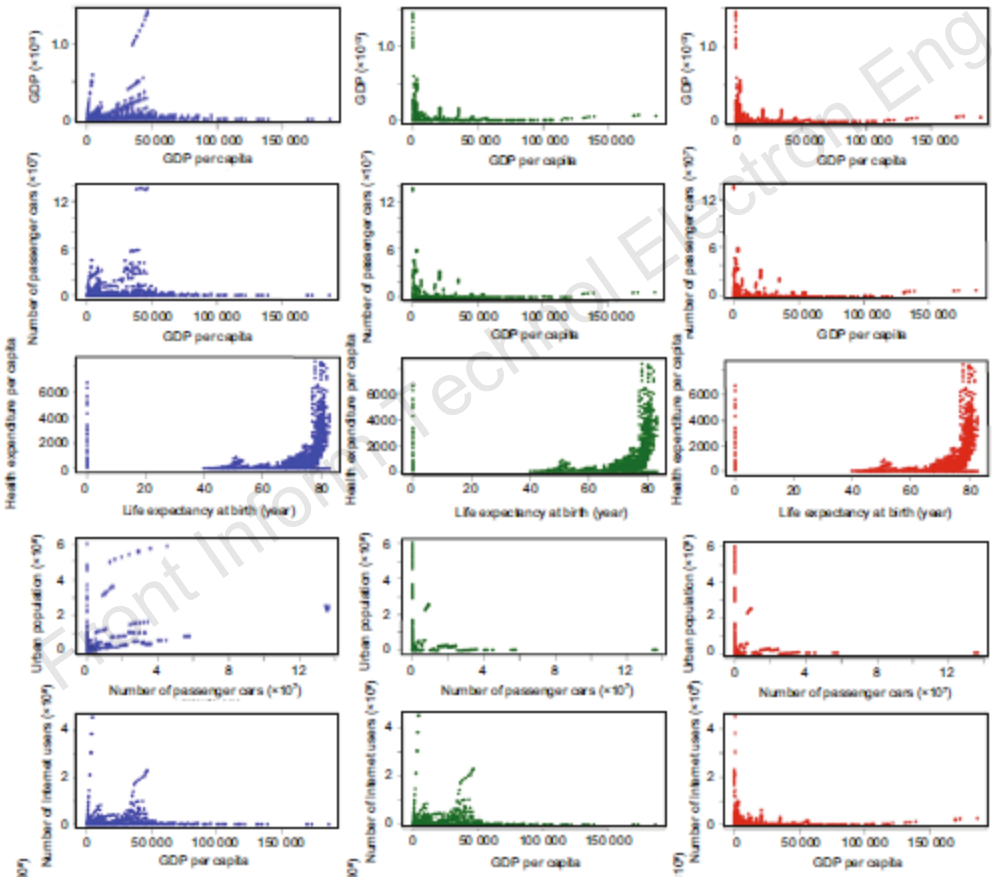

Fig. 6 Scatterplots depicting correlations between variables A and B by applying the two correlation coefficients. References to color refer to the online version of this figure

Conclusion and future scope

1. This work discusses a technique to improve a traditional correlation metric.

2. It defines a new general rank based correlation coefficient *g* which gives more intuitive correlation values among two variables. It is more accurate when the number of observations are large.

3. This metric is an improvement over the existing Spearman's rank correlation. It is more predictive and flexible in nature. It fares well on all five predictive metric criteria.

4. A lot of work can be done to refine this metric and making it more concrete like finding it's statistical significance.

5. The most important application of this technique will be in the field of prediction analysis. It can be used in many areas, such as online predictive systems, health care information systems, political prediction systems, and financial markets recommendation systems.