# Predictive General Rank Based Correlation Coefficient

## 1. CONCEPT BEHIND APPROACH

In order 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. These are the problems of dimensionality, outliers, coefficients showing fake correlations which do not exist etc. In this proposal the focus is on subsets of observations which have same multivariate features. There is a need to perform multivariate feature selection and identification of predictive set of metrics which best suit our purpose. The predictive metric should fulfil five main criteria,they are described in Figure 1

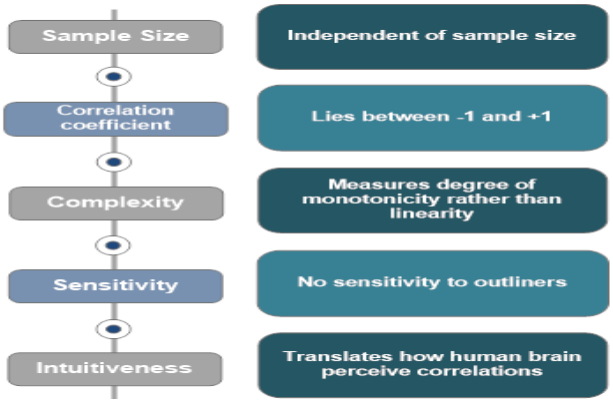


**Figure 1: Criteria for a Predictive Metric**

## 2. APPROACH AND UNIQUENESS

A predictive metric 'g' [General Rank Based Correlation Coefficient(GRCC)] is proposed which is in the form of multiple steps and fulfils all the above mentioned criteria. This metric finds correlations among two variables 'x' and 'y'.We assume that out of the two variables 'x' and 'y', 'x' is ordered. 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. 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. 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 reverse order of y in 'b'. Mathematically, g can be expressed as: $g = s * [1 - \frac{min(a,b)}{d}]$
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'. 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,

The value of the denominator 'd' can be selected in any of the ways given in Figure 2

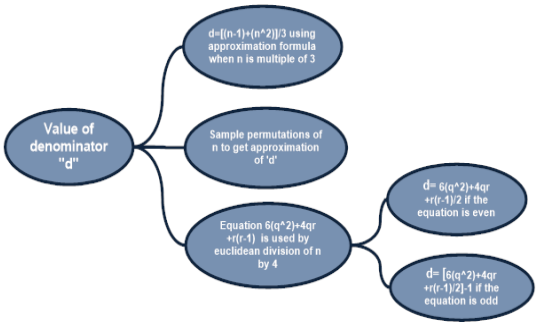


**Figure 2: Ways to Select Value of denominator d**

In order to test the proposed approach a data set was selected from `http://data.worldbank.org/indicator`. This data set was then tested on a data model created by using the general rank based correlation coefficient technique. The results are compared with the spearman's rank correlation to validate the results. The Spearman's rank correlation coefficient is given by: $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, Here $d_i = x_i - y_i$ [difference between ranks of observations] and lies between 1 and -1. The comparative analysis is given in table I. '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 dis-

**Table 1: Comparison of Spearman rank correlation and general rank correlation coefficient**

| A | B | No. of data points | Spearman's Rank Correlation Results | | | General Rank Correlation Coefficient | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | p value | S value | Rho($\rho$) | $d = 6q^2 + 4qr + \frac{r(r-1)}{2}$ | | $d = (n-6) + \frac{(n^2)}{3}$ | |
| | | | | | | g(c=1) | g(c=2) | g(c=1) | g(c=2) |
| GDP per capita | GDP | 2327 | <2.2e-16 | 900579181 | 0.4600 | 0.4302 | 0.3216 | 0.413 | 0.3092 |
| GDP per capita | Passenger cars | 2354 | <2.2e-16 | 1451827312 | 0.3321 | 1 | 1 | 1 | 1 |
| Life expectancy | Health exp. per capita | 2327 | <2.2e-16 | 858399753 | 0.538 | 0.9999 | 0.9984 | 0.9997 | 0.9842 |
| Passenger cars | Urban population | 2354 | <2.2e-16 | 1365948542 | 0.3717 | 0.748 | 0.989 | 0.703 | 0.942 |
| GDP per capita | Internet users | 2354 | <2.2e-16 | 1101037602 | 0.4935 | 0.7551 | 0.8821 | 0.8155 | 0.8122 |

tances. Figure 3 shows scatterplots depicting correlations between variables A and B.

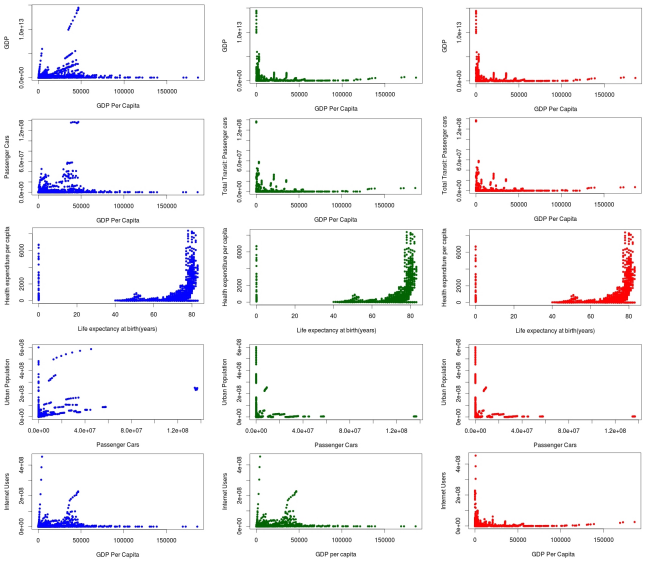


**Figure 3: Scatterplots Depicting Correlations Between Variables A and B**

On study of the values of the correlation coefficients obtained, it becomes clear that the 'g' is more intuitive than the traditional Spearman's rank correlation coefficient. 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. For instance there is a very high correlation between the car ownership and the GDP per capita which makes sense and the relatively low value of Spearman's rank correlation is not indicative of that.