PERFORMANCE STUDY OF LINFOOT'S INFORMATIONAL CORRELATION COEFFICIENT AND ITS MODIFICATION

G.L. Shevlyakov[1], N.V. Vasilevskiy[2]
[1,2]*Peter the Great St. Petersburg Polytechnic University*
[1]*Institute for Problems of Mechanical Engineering, Russian Academy of Sciences*
Saint Petersburg, RUSSIA
e-mail: [1]`Georgy.Shevlyakov@phmf.spbstu.ru`

Abstract

Performance of the informational correlation coefficient (Linfoot, 1957) is experimentally studied. To reduce the bias of estimation, a symmetric version of this correlation measure is proposed. This modified informational correlation coefficient outperforms Linfoot's correlation measure at the bivariate normal distribution on large samples.

1 Introduction

Pearson's correlation coefficient is a well-defined measure of the linear dependence between continuous random variables X and Y. This partially refers to closely related to it rank measures as the quadrant, Spearman and Kendall correlation coefficients. However, if one is interested either in processing discrete data or in revealing the possible nonlinear relationship between random variables, then difficulties may arise both in the implementation of those classical measures as well as in their interpretation.

In the literature, several proposals were made to solve these problems, for instance, Gebelein's (1941), Sarmanov's (1958) correlation coefficients, and the distance correlation coefficient of Szekely (2007).

In what follows, we focus on the informational measures of association between random variables [7]. Joe's dependence measure [4] exploits the concept of the relative entropy that measures the similarity of two random variables with the distributions $p(x)$ and $q(x)$ in the discrete case

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)}\,.$$

Silvey [8] uses the measure of dependence between two random variables defined by the ratio of their joint density and the product of their marginal densities $\varphi(x,y) = p(x,y)/[p(x)p(y)]$. The introduced measure is defined as follows: $\Delta = E[d(x)]$, where $d(x) = \int_{y:\,\varphi(x,y)>1}\big[p(y|x) - p(y)\big]\,dy$. Thus, it can be rewritten as

$$\Delta = \int\int_{(x,y):\,\varphi(x,y)>1}\big[p(x,y) - p(x)p(y)\big]\,dxdy.$$

Granger [2] introduces another measure of dependence

$$S_p = \frac{1}{2} \int \int \left[p(x,y)^{1/2} - [p(x)p(y)]^{1/2} \right]^2 \, dxdy.$$

Joe's measure of dependence is not symmetric, and Silvey's and Granger's measures are hard to compute. Mutual information ($I(X,Y)$) for any pair of discrete and continuous random variables X and Y is defined as follows

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \quad I(X,Y) = \int \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \, dxdy \, .$$

The informational correlation coefficient (ICC), firstly introduced by Linfoot [6], is defined as follows

$$\rho_{ICC}(X,Y) = \sqrt{1 - e^{-2I(X,Y)}} \, . \tag{1}$$

Note that ICC is equal to the classical Pearson's correlation coefficient at the bivariate normal distribution: $\rho_{ICC}(X,Y) = \rho$.

2 Problem Setting

In spite of the fact that ICC was introduced more than 60 years ago, its properties as a statistical measure of correlation have not yet been studied; it was not checked how well this measure estimates the correlation coefficient based on the sample of a given size. We are going to experimentally examine the following statistical properties of ICC: (i) unbiasedness, (ii) consistency, (iii) Monte Carlo performance on small ($N \leq 20$) and large samples, and (iv) robustness. Moreover, in order to improve the performance of ICC, namely, to reduce its bias, we propose and study a modified symmetric version of ICC denoted as $MICC$.

3 Monte Carlo Experiment

3.1 Description of the computational algorithm

All numerical experiments are performed using R language, especially its "entropy" library. The first problem is how to compute mutual information, which is used in (1). This is solved by applying a shrink-algorithm [3].

There exist several different algorithms of computing $I(X,Y)$; in our work, we choose the most precise one, not the fastest (for comparative analysis, see [3]). All experiments are performed at the standard bivariate normal distribution with density $f(x,y) = N(x,y;0,0,1.1,\rho)$.

The general algorithm can be described as follows:

1. Generate a sample of the fixed size: $N = 20, 60, 100, 1000, 10000$.
2. Extract x- and y-components from the sample, which are dependent random variables with the correlation coefficient ρ.

3. Construct the table of frequencies–the discrete analog of the joint distribution–we take a rectangle $[x_{min}, x_{max}] \times [y_{min}, y_{max}]$ on plane and divide it into $n_x \times n_y$ "bins" of equal size. Thus, the table of dimension n_x, n_y is built, each element of which is equal to the number of points in the corresponding bin.

4. Mutual information $I(X, Y)$ and ICC are computed using this table of frequencies.

This sequence is repeated 1000 times, allowing us to compute Monte Carlo estimates of the mean and variance of the correlation coefficient ρ: computations are performed for $\rho = 0, 0.1, 0.2, \ldots, 0.9, 1$; the number of bins is taken equal to 400. Typical results are exhibited in Fig. 1.

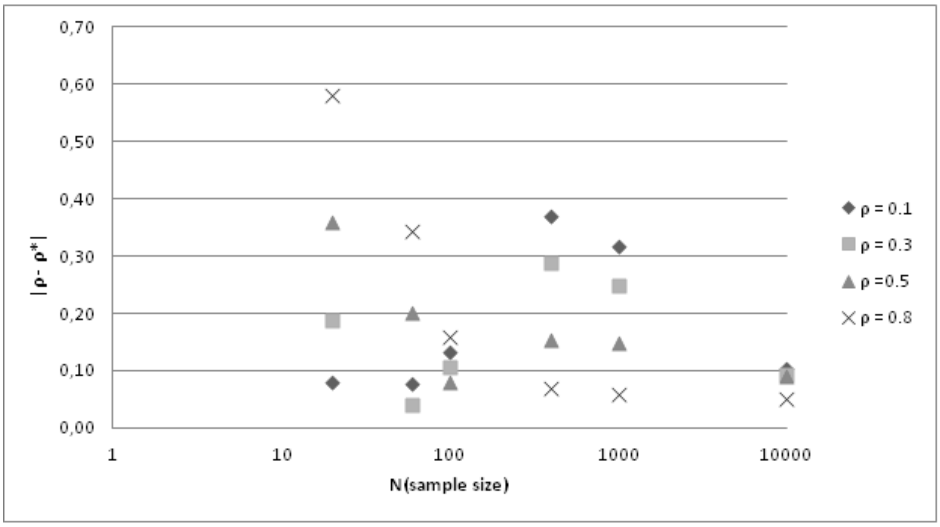

Figure 1: Monte-Carlo Biases of ICC

3.2 Monte Carlo results for *ICC*

1. From Fig. 1 it follows that estimation biases are considerably big (on small samples, they can even be greater than 0.5. Relatively small biases are observed only on large samples $N = 1000$ and$N = 10\,000$.

2. Satisfactory performance is observed in the case of a strong correlation—biases decrease with the growth of the sample size.

3. We may also add that the coefficient of variance is less than 0.2 for all examined combinations of (ρ, N).

4. *A remark on the choice of the number of bins.* The shrink-algorithm takes the table of frequencies as an input. It appeared that the algorithm performance depends on the relation N/K^2, where K is the linear dimension of the table. We observed that results are almost independent of the changes of K, as they

depend only on the coefficient $B = N/K^2$. For $\rho = 0.5$, the value $B = 7$ is optimal. Given a data sample, we can choose an appropriate value of K, which is a variable in our algorithm.

4 Main Result: A Symmetric Modification of *ICC*

Mutual entropy, also known as the Kullback-Leibler distance, has a serious disadvantage — it is not symmetric, i.e., $D_K L(p||q) \neq D_K L(q||p)$. Thus, the Kullback-Leibler divergence is used [5]:

$$Div(p||q) = D_{KL}(p||q) + D_{KL}(q||p). \tag{2}$$

Analogously, a symmetric version of mutual information can be introduced

$$J(x, y) = I(x, y) + I^*(x, y)$$
$$= \int\int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dxdy + \int\int f(x)f(y) \log \frac{f(x)f(y)}{f(x, y)} dxdy\,.$$

Our idea is to repeat Linfoot's derivation of formula (1), replacing the mutual information $I(X, Y)$ with its symmetric version $J(X, Y)$. In this case, the following result holds.

Theorem 1. *A modified symmetric analog of the Linfoot's informational correlation coefficient (1) called as the modified informational correlation coefficient (MICC) is given by:*

$$\rho_{MICC} = \sqrt{1 - \frac{2}{W(2e^{2(J+1)})}}, \tag{3}$$

where $W(x)$ is the Lambert function [1], inverse w.r.t. xe^x. In particular, $\rho_{MICC} = \rho$ at the standard bivariate normal distribution.

The results of comparison of these two correlation measures are exhibited in Fig. 2–Fig. 4: from them it follows that *MICC* outperforms *ICC* on all examined combinations of sample sizes and correlation coefficients. The observed improvement is more considerable on small samples and low values of the correlation coefficient — just in the most difficult cases for *ICC*.

5 Conclusions

1. The statistical performance of the Linfoot's informational correlation coefficient is studied: considerable biases of estimation are observed.

2. To reduce the biases of *ICC*, a modified symmetric version of it, namely *MICC*, is proposed, which proved to provide much lesser estimation biases as compared to the original one.

3. The proposed modified informational correlation coefficient *MICC* is recommended for processing Big Data.

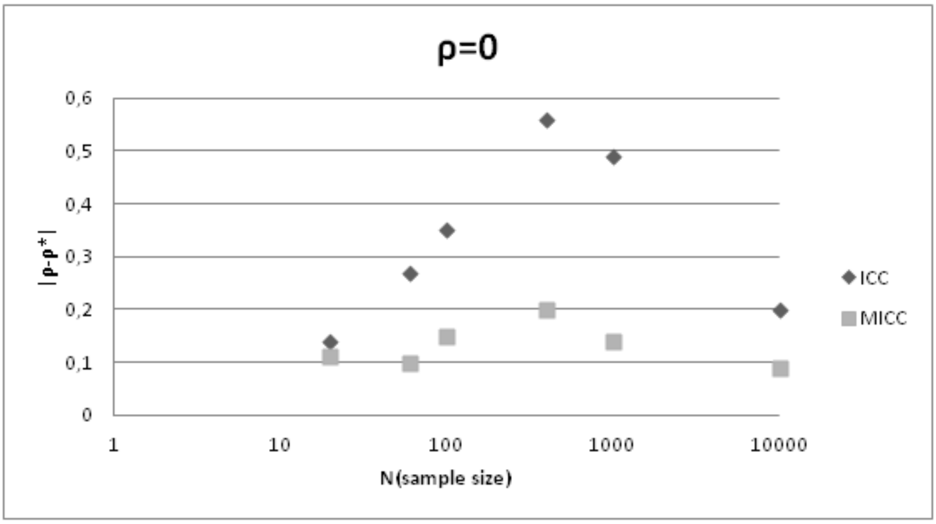

Figure 2: Monte-Carlo Biases of *ICC* and *MICC* at $\rho = 0$

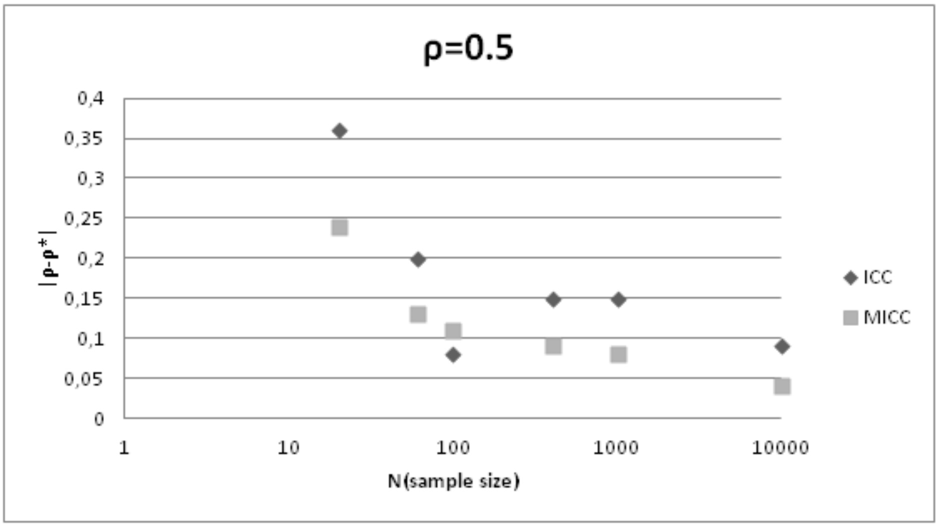

Figure 3: Monte-Carlo Biases of *ICC* and *MICC* at $\rho = 0.5$

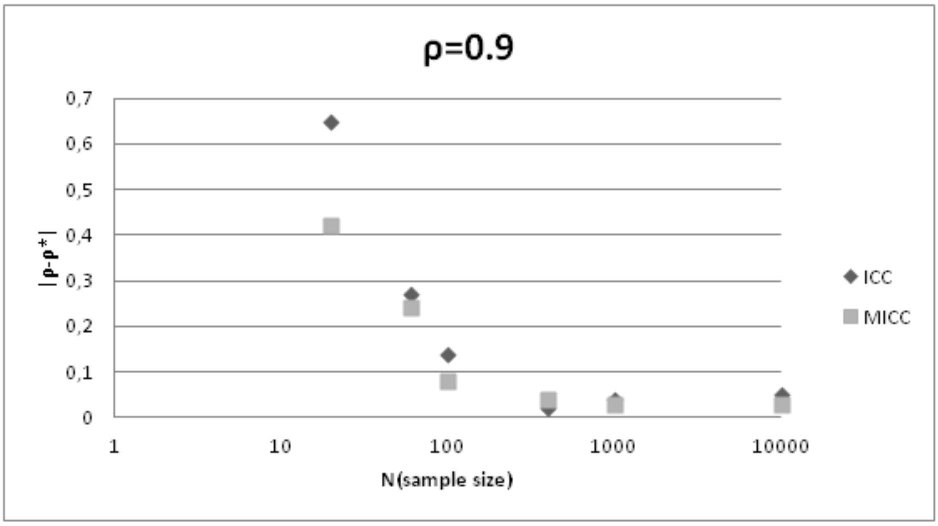

Figure 4: Monte-Carlo Biases of *ICC* and *MICC* at $\rho = 0.9$

References

[1] Corless R., Gonnet G., Hare D., Jeffrey D., Knuth, Donald. (1996). On the Lambert W Function. *Advances in Computational Mathematics*. (Berlin, New York: Springer-Verlag). Vol. **5**, pp. 329-359.

[2] Granger C.W., Maasoumi E., Racine J. (2004). A Dependence Metric for Possibly Nonlinear Processes. *Journal of Time Series Analysis*. Vol. **25**, pp. 649-669.

[3] Hausser J., Strimmer K. (2010). Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *Journal of Machine Learning Research*. Vol. **10**, pp. 1469-1484.

[4] Joe H. (1989). Relative Entropy Measures of Multivariate Dependence. *Journal of the American Statistical Association*. Vol. **84**, pp. 157-164.

[5] Kullback S. (1959). *Information Theory and Statistics*. John Wiley.

[6] Linfoot E.H. (1957). An Informational Measure of Correlation. *Information and Control*. Vol. **1**, pp. 85-89.

[7] Shannon C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. Vol. **27**, pp. 379-423, 623-656.

[8] Silvey S.D. (1964). A Measure of Association. *The Annals of Mathematical Statistics*. Vol. **35**, pp. 1157-1166.