Comment on "Detecting Novel Associations in Large Data Sets"

Malka Gorfine[1]

Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology,
Technion City, Haifa 32000, Israel
gorfinm@ie.technion.ac.il

Ruth Heller

Department of Statistics and Operational Research, Tel-Aviv University, Israel
ruheller@post.tau.ac.il

Yair Heller

heller.yair@gmail.com

Abstract: Reshef *et al.* presented a novel measure of dependence - the maximal information coefficient (MIC) aimed to capture a wide range of associations between pairs of variables, and a statistical test for independence based on MIC. They defined a concept of equitability and claim that non-equitable methods are less practical for data exploration. By simple power comparisons, we show that this conclusion is wrong.

As pointed out by Reshef *et al.* (Research Article, 17 Dec 2011, p. 1518), it is often the case that the pairwise relationship between many variables is simultaneously explored. In statistics, this exploration is formalized in a multiple hypothesis testing framework, where the null hypothesis of statistical independence is examined for every pair of variables. Then, the p-values of the tests serve as a basis for generating final conclusions. Specifically, the pairs of variables are ordered by their p-values (or the adjusted p-values after correcting for multiple testing) in increasing order, and the pairs with the lowest p-values will be further studied. Reshef *et al.*

[1] To whom correspondence should be addressed

recommended ranking the pairs based on MIC, which in this case is equivalent to ranking based on the p-values of the MIC tests, for fixed sample size.

A highly important concept in hypotheses testing is the *power* of a statistical test: the probability that the test will correctly lead to the rejection of the null hypothesis when the null hypothesis is false. The power of a test usually depends on various factors, such as the sample size. Two desirable properties of a test are the following. First, that the power increases as the sample size increases, ie that the test is *consistent* against all alternatives. Second, that the test achieves good power for finite sample sizes. Therefore, when comparing various statistical tests for the same problem of hypotheses testing, it is important to compare their power under different sample sizes and different configurations under which the null hypothesis is false.

Reshef *et al.* contrasted MIC with other dependency measures under fairly large sample sizes, such as 250, 500 or 1000. They compared how the values of the dependency measures change across different functional relationships, when the same level of noise was added to these functional relationships (see Fig. 2 of their paper), and they ignored the power of the tests based on the measures. They concluded that since the other considered methods are not equitable even in the basic case of functional relationships (they showed a strong preference for some type of functions even at identical noise levels) these methods are less practical for data exploration compared to MIC. In the following simulation results we demonstrate that under practical sample sizes, such as 50, and functional relationships at identical noise level, the power of the MIC test varies dramatically between the various relationships. That is, under practical sample sizes, MIC tends to have strong preference for some type of functions even at identical noise levels. Moreover, it will be shown that under various settings, other tests which enjoy the advantage of proven consistency, such as the distance correlation of Szekely *et al. (3)* (which was considered by Reshef *et al.*) and HHG of Heller *et al. (1)*, are more powerful than the test based on MIC and thus are preferable over MIC.

Tables 1-2 present the results of a small simulation study comparing the power functions of three tests: the distance correlation (dCor) *(3)*, MIC, and a new test, HHG, recently proposed in *(1)*, while focusing on practical sample sizes. For each configuration considered, 1000 dataset were generated and each of the three tests was performed at a significance level of $\alpha = 0.05$.

The reported empirical power is the proportion of tests concluded with rejection of the null hypothesis. Table 1 mimics the functional relationships studied by Reshef *et al.* (SOM of *(2)*, Section 4.3) with and without noise. It is evident that under the noiseless setting, MIC substantially outperforms dCor and slightly outperforms HHG in 8 and 5 configurations, respectively; whereas under a noise of about 0.64 dCor and HHG outperform MIC in 11 and 10 functions, respectively. The configurations of Table 2 are unusual bivariate relation presented and plotted in the wikipedia.org page on Pearson correlation. The 4 independent clouds represent the case where the null hypothesis is true, and thus the empirical power is expected to be the level of the test, 0.05. It is evident that the HHG test outperforms the other tests at any sample size and for any false null distribution considered except for the W shape and a sample size of $n = 30$. With $n = 220$ and the diamond relation, the respective values of the empirical power of dCor and MIC (standard error) are: 0.799(0.0127) and 0.08(0.0086); and under the circle relation the respective values are: 1(0), 1(0). With $n = 600$ and diamond relation the empirical power of MIC equals 0.209(0.0129). It should be noted that to the best of our understanding the proof of consistency of MIC does not hold for this type of data and indeed the performance is dismal.

Summarizing, it is evident that (i) under the (unrealistic) noiseless functional settings, MIC tends to slightly outperform HHG and significantly outperforms dCor; (ii) under the majority of the noisy functionals and non-functional settings, the HHG and dCor tests hold very large power advantages over the MIC test, under practical sample sizes; (iii) results obtained from practical sample sizes (30, 50 or 100) could be markedly different from the results obtained from larger samples (250, 500 or 1000) as considered in Reshef *et al.* These results clearly contradict Reshef *et al.*'s conclusion that a method such as dCor is less practical for data exploration. Additionally, consistency for certain type of dependencies is proven for MIC which is not feasible to calculate, and it is not proven for the approximation actually being used. To the best of our knowledge there are very few consistent scores that are of simple form, and here we focused the power comparison on two such consistent tests, dCor and HHG, that are also applicable for testing for independence between two random vectors of arbitrary dimensions while MIC is only applicable for one dimensional variables.

References

1. R. Heller, Y. Heller, M. Gorfine, A consistent multivariate test association based on ranks of distances. *Front for the Mathematics ArXiv*, under **Statistics**, arXiv:1201.3522v1, (2012).

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

3. G. Szekely, M. Rizzo, N. Bakirov, Measuring and testing independence by correlation of distances. *The Annals of Statistics,* 35: 2769-2794, (2007).

Table 1: *Simulation results: The empirical power (SE) for a test at level 0.05, based on 1000 datasets for each configuration of joint distribution and sample size n=50. The distributions are those of Table S3 at Reshef et al. ((2), SOM). The tests compared: dCor (based on the 'energy' R package with 1000 repetitions per dataset), MIC (with parameters and p-values provided by the MINE website exploredata.net); HHG (the code can be provided by the authors upon request).*

	Noise equals 0			Noise is approximately 0.64		
Function Name	dCor	MIC	HHG	dCor	MIC	HHG
Linear+Periodic, Low Freq	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.938(0.0076)	0.987(0.0036)
Linear+Periodic, Medium freq	0.545(0.0158)	0.941(0.0075)	0.996(0.0020)	0.221(0.0131)	0.206(0.0128)	0.175(0.0120)
Linear+Periodic, High Freq	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.998(0.0014)	0.842(0.0115)	0.972(0.0052)
Linear+Periodic, High Freq 2	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.998(0.0014)	0.869(0.0167)	0.967(0.0057)
Non-Fourier Freq [Low] Cosine	0.126(0.0105)	1.000(0.0000)	0.997(0.0017)	0.068(0.0080)	0.293(0.0144)	0.106(0.0097)
Cosine, High Freq	0.057(0.0073)	0.328(0.0148)	0.281(0.0142)	0.050(0.0069)	0.083(0.0087)	0.063(0.0077)
Cubic	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.681(0.0147)	0.596(0.0155)	0.733(0.0140)
Cubic, Y-stretched	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.697(0.0145)	0.650(0.0151)	0.747(0.0138)
L-shape	0.999(0.0000)	1.000(0.0000)	1.000(0.0000)	0.112(0.0100)	0.040(0.0062)	0.117(0.0102)
Exponential [2^x]	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.944(0.0073)	0.626(0.0153)	0.892(0.0098)
Exponential [10^x]	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.471(0.0158)	0.166(0.0118)	0.573(0.0156)
Line	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.998(0.0014)	0.895(0.0097)	0.980(0.0044)
Parabola	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.589(0.0156)	0.471(0.0158)	0.817(0.0122)
Random	0.059(0.0075)	0.047(0.0067)	0.053(0.0071)	0.056(0.0073)	0.052(0.0070)	0.040(0.0062)
Non-Fourier Freq [Low] Sine	0.091(0.0091)	0.999(0.0010)	0.853(0.0112)	0.060(0.0075)	0.233(0.0134)	0.081(0.0086)
Sine, Low Freq	0.318(0.0147)	1.000(0.0000)	0.972(0.0052)	0.128(0.0106)	0.229(0.0133)	0.113(0.0100)
Sine, High Freq	0.108(0.0098)	0.296(0.0144)	0.235(0.0134)	0.084(0.0088)	0.073(0.0082)	0.061(0.0076)
Sigmoid	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.998(0.0014)	0.940(0.0075)	0.948(0.0070)
Varying Freq [Medium] Cosine	0.316(0.0147)	0.988(0.0034)	0.831(0.0119)	0.131(0.0107)	0.213(0.0129)	0.121(0.0103)
Varying Freq [Medium] Sine	0.238(0.0135)	0.883(0.0102)	0.631(0.0153)	0.117(0.0102)	0.129(0.0106)	0.098(0.0094)
Spike	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.645(0.0151)	0.261(0.0139)	0.591(0.0156)
Lopsided L-shaped	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)	0.101(0.0095)	0.061(0.0076)	0.102(0.0096)

Table 2: *Simulation results: The empirical power (SE) for a test at level 0.05, based on 1000 datasets for each configuration of joint distribution and sample size n. The distributions are various unusual bivariate relation presented in the wikipedia.org page on Pearson correlation. The tests compared: dCor (based on the 'energy' R package with 1000 repetitions per dataset), MIC (with parameters and p-values provided by the MINE website exploredata.net), and HHG (the code can be provided by the authors upon request).*

n	Distribution	dCor	MIC	HHG
30	W	0.000(0.0000)	0.887(0.0100)	0.727(0.0141)
	Diamond	0.029(0.0053)	0.084(0.0088)	0.284(0.0143)
	Parabola	0.470(0.0158)	0.734(0.0140)	0.899(0.0095)
	Hyper-parabola	0.189(0.0124)	0.144(0.0111)	0.997(0.0017)
	Circle	0.000(0.0000)	0.221(0.0131)	0.472(0.0158)
	4 independent clouds	0.041(0.0063)	0.042(0.0063)	0.050(0.0068)
50	W	0.815(0.0123)	1.000(0.000)	1.000(0.0000)
	Diamond	0.049(0.0068)	0.087(0.0089)	0.662(0.0149)
	Parabola	0.975(0.0049)	0.906(0.0092)	0.998(0.0014)
	Hyper-parabola	0.301(0.0145)	0.366(0.0152)	1.000(0.0000)
	Circle	0.000(0.0000)	0.359(0.0152)	0.993(0.0026)
	4 independent clouds	0.046(0.0066)	0.045(0.0066)	0.050(0.0068)
100	W	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)
	Diamond	0.155(0.0114)	0.072(0.0082)	0.990(0.0031)
	Parabola	1.000(0.0000)	1.000(0.0000)	1.000(0.0000)
	Hyper-parabola	0.951(0.0068)	1.000(0.0000)	1.000(0.0000)
	Circle	0.000(0.0000)	1.000(0.0000)	1.000(0.0000)
	4 independent clouds	0.052(0.0070)	0.034(0.0057)	0.052(0.0070)