

COMMENT ON “DETECTING NOVEL ASSOCIATIONS  
IN LARGE DATA SETS” BY RESHEF ET AL,  
SCIENCE DEC 16, 2011

NOAH SIMON AND ROBERT TIBSHIRANI

The proposal of Reshef et. al. (“MIC”) is an interesting new approach for discovering non-linear dependencies among pairs of measurements in exploratory data mining. However, it has a potentially serious drawback. The authors laud the fact that MIC has no preference for some alternatives over others, but as the authors know, there is no *free lunch in Statistics*: tests which strive to have high power against all alternatives can have low power in many important situations.

To investigate this, we ran simulations to compare the power of MIC to that of standard Pearson correlation and distance correlation (`dcor`) Székely & Rizzo (2009). We simulated pairs of variables with different relationships (most of which were considered by the Reshef et. al.), but with varying levels of noise added. To determine proper cutoffs for testing the independence hypothesis, we simulated independent data with the appropriate marginals. As one can see from the Figure, MIC has lower power than `dcor`, in every case except the somewhat pathological high-frequency sine wave. MIC is sometimes less powerful than Pearson correlation as well, the linear case being particularly worrisome.

This set of dependencies is by no means exhaustive, however it suggests that MIC has serious power deficiencies, and hence when it is used for large-scale exploratory analysis it will produce too many false positives. The “equitability” property of MIC is not very useful, if it has low power.

We believe that the recently proposed distance correlation measure of Székely & Rizzo (2009) is a more powerful technique that is simple, easy to compute and should be considered for general use.

A full R language script for our analysis appears in  
<http://www-stat.stanford.edu/~tibs/reshef/script.R>

---

Depts. of Statistics and Health Research and Policy, Stanford University. Stanford, CA. 94305; 650-723-5989.

## REFERENCES

- Székely, G. J. & Rizzo, M. (2009), 'Brownian distance covariance',  
*Annals of Applied Statistics* pp. 1233–1303.

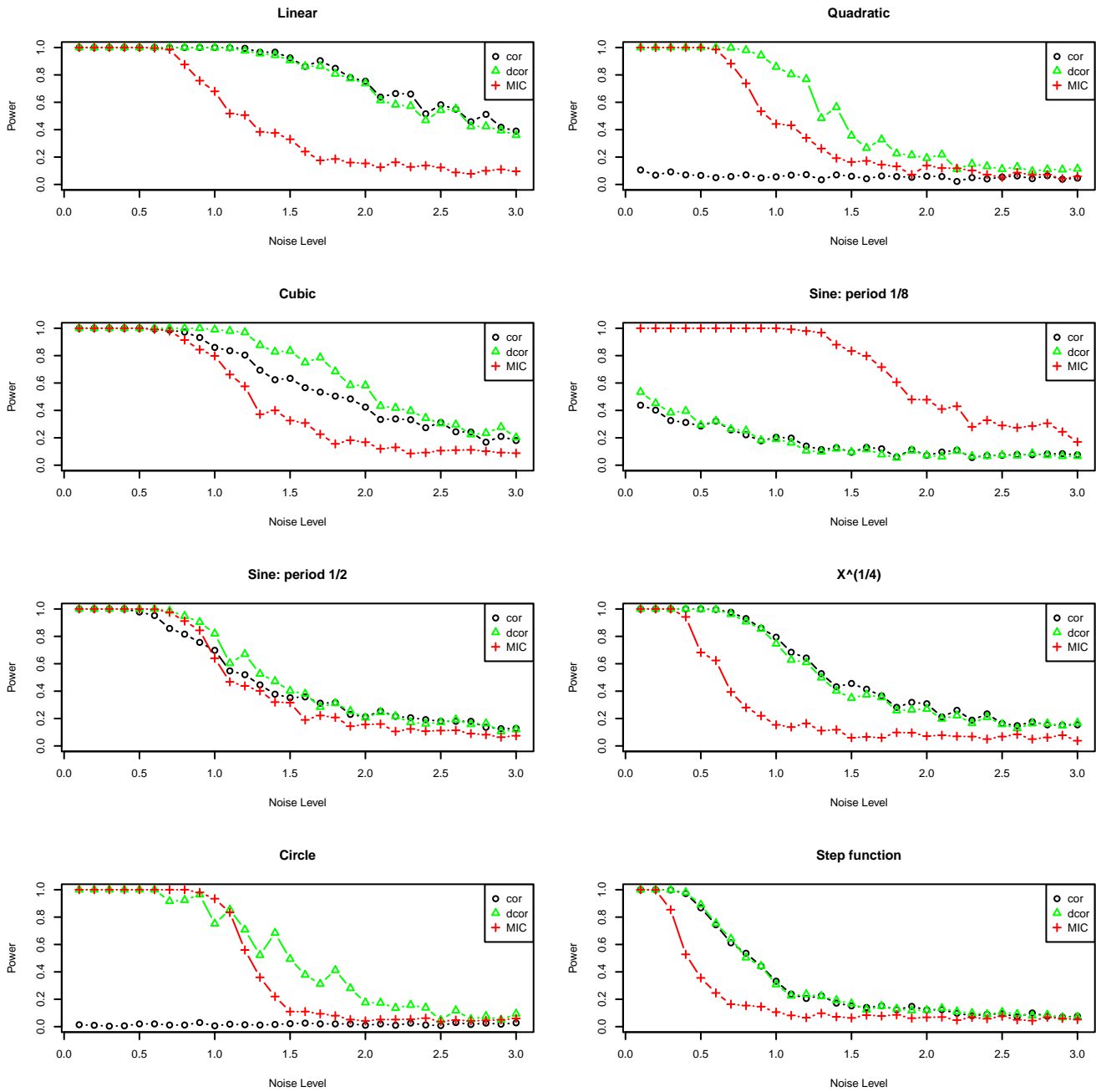


FIGURE 1. Power of Pearson correlation (cor), distance correlation (dcor) and MIC as a function of the level of noise added, in eight different scenarios. The power was estimated via 500 simulations. MIC as lower power than dcor in every case except the somewhat pathological high-frequency sine wave.