
# AN EVALUATION OF SOME SIMPLE MEASURES FOR DETECTING NON-LINEAR RELATIONSHIPS BETWEEN VARIABLES

**David Cornforth**

**Working Paper Series, Number 6, March 2014**

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

air

**ABOUT THIS SERIES**

*The Applied Informatics Research (AIR) Group is a cross-disciplinary, multi-institutional collaboration based in the School of Design, Communication & Information Technology at the University of Newcastle, Australia. The principal aim of the AIR Group's Working Paper Series is to disseminate the research and/or technical output of the group in an easily accessible format. The content of Working Papers generally falls into one of the following categories:*

- *Preliminary findings or results, the release of which is intended to stimulate debate and/or discussion to assist in the further development of the research*
- *Technical reports associated with applied research that may be written in a less academic style than usually published in academic journals*
- *Extended versions of published works, often containing additional implementation/ application detail, figures and tables.*

*The opinions or conclusions expressed in the Working Paper Series are those of the authors and do not necessarily reflect the views of the AIR Group as a whole.*

# AN EVALUATION OF SOME SIMPLE MEASURES FOR DETECTING NON-LINEAR RELATIONSHIPS BETWEEN VARIABLES

**David Cornforth**
School of Design Communications and IT
University of Newcastle
Callaghan 2308, NSW, Australia

David.Cornforth "at" newcastle.edu.au
http://silverbullet.newcastle.edu.au/davidcornforth/

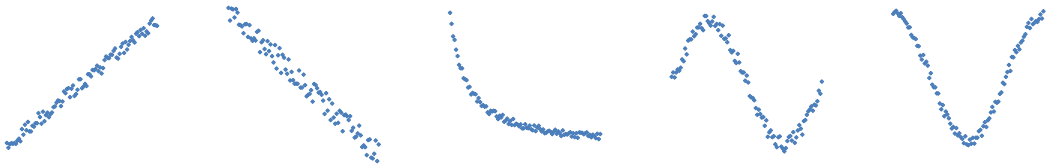
**ABSTRACT**

Since the introduction of simple measures of linear relationship such as Pearson's Correlation Coefficient, measures have been sought that will also describe non-linear relationships that may exist between a pair of variables. Currently there are a number of such methods, encompassing a range of sophistication and involving a range of computational effort. This work reports on some experiments with a computationally simple measure that operates using a division of input space into regularly spaced cells. The Distribution Area Ratio Correlation Coefficient (DARCC) compares the distribution of cells containing *k* points with a theoretical distribution. The method is described then evaluated by comparing the resulting correlation coefficient with the magnitude of added noise. Results show a good agreement between noise and DARCC for several synthesised datasets. The measure is also evaluated on some real datasets. DARCC is computationally very simple and has potential for datasets with a large number of variables where speed is important.

## 1 Introduction

Measures of correlation or association between variables are important in order to detect whether variables are related. Such measures have wide application. There are many measures including Pearson, Spearman, Mutual Information, Principal Curves, Maximal Correlation, and Maximal Information Coefficient (MIC). All these methods have advantages and disadvantages. In this paper I review some of these, and then focus on a method that relies on the distribution of points falling into cells of a regular grid in the X,Y space formed by the two variables being assessed. This measure may sacrifice some desirable qualities of more exact measures. However, the method is computationally efficient and so may have potential for implantations where speed is critical, such as the screening of large numbers of variables.

A correlation measure is an indication of how two variables are related. Such measures are widely used to assess the dependence between two variables. The most well-known measures of correlation are Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient. However, these measures only apply to linear functions. If the variables are related in a nonlinear way, other measures must be sought. Figure 1 shows several sets of 2-dimensional data that are strongly related. Figure 1 (a) and (b) show a linear relationship that is close to perfect, as indicated by a Pearson coefficient close to ±1. The data shown in Figure 1 (c) are not related in a linear fashion but the Pearson coefficient provides a high value in spite of this, because the distribution of points places more emphasis on the relatively linear part of the curve towards the right hand side. The data in (d) are even more nonlinear, yet the Pearson coefficient is high, because the frequency and phase just happen to align a major proportion of the points in a diagonal fashion from top left to bottom right. This alignment cannot be expected to occur in such a fortuitous manner in real data. The data in (e) are the same, except for a phase shift, yet the Pearson coefficient does not indicate any relationship. These examples show the limitations of measures such as Pearson, and suggest that a measure that could detect nonlinear relationships in a reliable way would be of great use in many applications. The following section will describe some existing measures.



(a) 0.994 (b) -0.987 (c) -0.83 (d) -0.672 (e) -0.012

Figure 1: Datasets showing simple relations with their Pearson correlation coefficient.

## 2 Existing Correlation Measures for Nonlinear Functions

This section describes some existing measures that provide a result indicating the degree of relationship between two variables.

### 2.1 Mutual information

Mutual Information (MI) estimates information that is common to both X and Y. An efficient method exists for calculating MI using fixed divisions, or bins, of the ordinates X and Y of a data set (Kraskov, 2004). Here the X and Y ordinates are divided into bins to provide a 2-dimensional grid or histogram. The area of intersection between a horizontal and vertical bin will be referred to here as a cell. MI is estimated directly from counting the number of points that fall into each cell:

$$MI(X,Y) = \sum_{ij} p(i,j) log \frac{p(i,j)}{p_x(i)p_y(j)}$$

where *i* and *j* are indices of the bins in the X and Y direction. The probabilities are estimated from a count of points falling into the relevant area: $p(i,j)$ from points falling into cell $(i,j)$, $p_x(i)$ from points falling into the *ith* bin of X and $p_y(j)$ from points falling into the *jth* bin of Y, each divided by the total number of points. If the variables are not related, the joint probability density is equal to the product of the marginal probability densities and MI is zero. If they are related then MI tends towards infinity (Moon et al., 1995). In order to use this as a correlation measure, it would be desirable to provide a measure, based on MI, but having a bounded range in [0,1].

Any method using a regular grid may be sensitive to where the bin boundaries are placed, as small shifts in boundary can move points into adjacent bins. Kernel density estimation extends this method by replacing each point by a kernel function, so that each point also makes a small contribution to neighbouring bins. This is superior to the histogram because of higher resolution, and because it is less sensitive to the choice of histogram bin boundaries (Moon et al., 1995). The calculation of kernel functions make the estimation more computationally complex.

Another refinement of MI makes use of the ranks of the variates instead of their value (Wang et al., 2005). In this measure, samples in X and Y are individually sorted into bins with an equal number of ordinates falling into each bin. Then the points are sorted according to the order of their ordinates.

Maximal Information Coefficient (MIC) uses an irregular division of the X, Y space which is optimised in order to provide the maximum measure of Mutual Information (Reshev et al., 2011). The authors provide a large number of simulations to support the accuracy of this method. This extra refinement introduces more computational effort, suggesting that there is a trade-off between speed and accuracy in these methods. The method presented in section 4 of this paper explores the high speed end of this continuum.

### 2.2 Entropy

A related measure (Wang et al., 2005) provides an estimate of correlation coefficient based on entropy:

$$CC = 2 + \sum_{k=1}^{b^2} \frac{n_k}{N} log_b \frac{n_k}{N}$$

where $b$ is the number of bins in each direction, $b^2$ is the total number of cells, $n_k$ is the number of points falling into cell $k$, and $N$ is the total number of points. Wang et al. (2005) argue that this entropy measure falls into a closed interval [0,1] while MI does not.

Another improvement in accuracy can be obtained by making the bins of varying size (Darbellay & Vajda, 1999). However, this involves another stage of estimation of the optimum bin boundaries, and increases the computational effort.

### 2.3 Maximal Correlation

A different approach is the maximal correlation, that proposes a transformation of both variables in X and Y to $f(x)$ and $g(y)$. These functions are chosen such that the linear correlation coefficient $r$ is maximised:

$$MC(x,y) = sup_{f,g} R(f(x), g(y))$$

(Rényi, 1959). The functions $f(x)$ and $g(y)$ can be found using an iterative procedure which uses Alternating Conditional Expectations (ACE). This algorithm has a guaranteed convergence (Breiman & Friedman, 1985), however it also has the computational load required by the iteration.

### 2.4 Principal Curves

A principal curve is a smooth curve that passes through a dataset, minimising errors while using some constraints and some measure of distance to define errors. The simplest case is a least squared regression line that can summarise a dataset (X, Y) using a straight line which minimises squared errors in Y. A principal component line is similar, but minimises squared errors in both X and Y. A principal curve minimises squared errors in X and Y, but is subject to smoothing constraints and uses the orthogonal distance from the curve to data points (Hastie & Stuetzle, 1989). The subset of points used for estimation of any part of the curve is selected using a clustering algorithm.

Tibshirani (1992) draws attention to possible bias inherent in the method of Hastie & Stuetzle (1989) and proposes an improved algorithm that makes use of the Expectation Minimisation (EM) algorithm. However this work presents no evidence that this works better than the former method.

Dependence measures based on principal curves include the covariance and the correlation along the curve (Delicado & Smrekar, 2009). These measures conform to a list of desirable properties articulated by Renyi (1959).

### 2.5 Distance Correlation

The Brownian Distance Correlation (Székely & Rizzo, 2009) relies on measuring the Euclidean distance between all pairs of points then subtracting arithmetic means in both directions of the resulting matrices. These matrices are then multiplied and summed, then normalised to provide a correlation measure.

## 3 Desirable Properties of a Simple Correlation Measure

Although the measures briefly reviewed above provide useful correlation measures, the intention of this work is to focus on simple measures that are computationally efficient. In consequence, there

must be some trade-off between accuracy and computational speed. This suggests a modest list of desirable properties of such a measure:

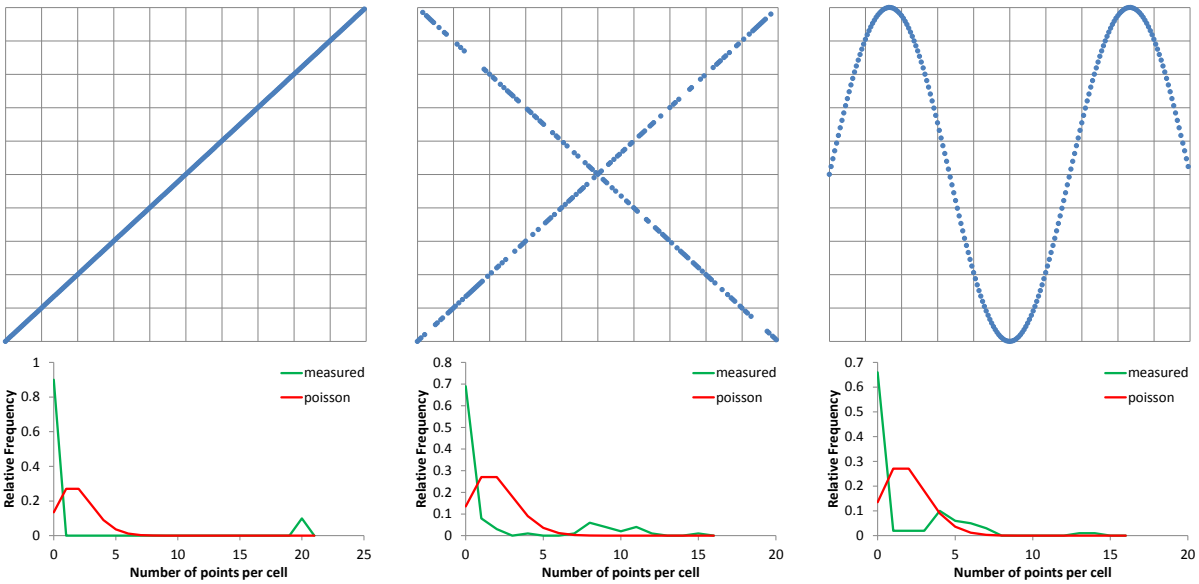
1. The measure must be bounded by [0,1]. The sign becomes irrelevant for arbitrary functions, since the idea of a negative slope is only feasible for a limited number of relationships.
2. The measure should exhibit a linear response to added noise. A noiseless function should achieve a correlation score of 1.0, while a function with 50% added noise should achieve a correlation score of 0.5 and so on.
3. The correlation measure should be invariant over the type of function, within reason. This means that, for example, a noiseless linear, cubic or sine function should all achieve a correlation score of 1.0. The idea of "within reason" means that the complexity of the function will have some requirement for sufficient sample size to describe that function adequately.

In pursuit of such a measure, this work will consider only measures derived using a fixed regular grid. This will remove the need for optimised grid spacing. The distribution of points falling into regular grid spacing will be examined, for a variety of functions. For uniform random noise (representing a bivariate dataset with no correlation) the theoretical distribution of points within a cell can be modelled as a Poisson distribution, where the probability of a cell containing *k* points is *P(X=k):*

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda$ is the expected value of *x*. If there are *n* points and *c* cells, the expected value $\lambda$ is *n/c*. Detection of a relationship between the two variables may be possible by examining any departure from this distribution.

Figure 2 shows a variety of datasets generated from functions with their corresponding distribution, shown with a green line. The linear function (a) has a clear bimodal distribution, with all cells either containing a similar number of points, or no points. As there are 200 points, and they fall into 10 cells on the diagonal, the distribution will consist of 10 cells containing 20 points each, and the remaining 90 cells containing zero points. The cross function (b) has a less pronounced bimodal distribution, but the distribution can be seen to be far removed from the Poisson. The sine function is a somewhat extreme example, and appears the most problematic of all, with a distribution closer to the Poisson than the other two. Initial examination of these distributions would suggest that the development of a measure, based on the distribution of points in cells, which would satisfy the desirable criteria above, would be very difficult.




a) A linear function b) a cross function c) a sin function

Figure 2. Several datasets of 200 points, generated from 3 different functions, overlaid with a regular grid. Corresponding distributions are shown below (green line), with the Poisson distribution (red line) for comparison.

## 4 Implementation of Grid-Based Correlation Measures

An examination of the distributions of generated datasets in Figure 2 shows three main sections, the large peak on the left, the tail of the distribution, and the central portion containing the peak of the Poisson distribution. In spite of the differences, these distributions all have one thing in common: a bimodal distribution.

The large peak on the left is the number of empty cells, which is much higher when the variables are related by a function than for a random distribution, and this is true even for a complex function. However, this number does not have an upper bound that is the same for all functions, as when measuring a given data set it is not known which function is represented therein. However, this figure alone has the potential to provide a "go / no-go" test on whether there is a relationship between the variables or not. Because of the bounding issue, it is difficult to adapt this to provide a continuous correlation measure between 0 and 1 that will be consistent for different functions.

The tail of the distribution has a very different shape for different functions, and also does not have an upper bound, so is not pursued here.

The central portion of the distribution has both an upper and lower bound. The upper bound is the value of the Poisson distribution, while the lower bound is zero. Between limits on the horizontal axis, the area under the curve is similar for different functions, and this will be developed as a correlation measure. This measure, the Distribution Area Ratio Correlation coefficient (DARCC) is based on the area under the measured curve, divided by the area under the Poisson curve, with the limits $0 < k \leq \lambda$. Areas are estimated by numerical integration based on integer values of $k$.

## 5 Experimental Methods

In order to investigate the characteristics of this measure, the algorithm was used on a variety of functions, with the number of bins varied within the bounds mentioned previously, that is, *k* must be ≥ 1. This leads to the maximum number of bins that are shown in Table 1.

Table 1. Number of bins divisions used in experiments.

| points | min bins | max bins |
|---|---|---|
| 500 | 5 | 22 |
| 250 | 5 | 15 |

Four functions were investigated:

a) Linear: $y = x$
b) Squared: $y = x^2$
c) Cross: half of data follows $y = x$, the other half $y = 1 - x$
d) Sine: $y = sin(2\pi fx)$, where $f = 2.7$

The functions were evaluated for values of x equally distributed in the range [0, 1]. Generated values of y were normalised to fit into the range [0, 1]. Noise was varied between 0 and 1 in steps of 0.1. Noise was added to all functions to create a new variable with a given amount of uniform random noise using:

$$y^* = vU(0,1) + (1 - v)y$$

where *y** is the new variable, *v* is the noise level in the range [0, 1], and *U(0, 1)* is a uniform random number if the range [0, 1]. After adding noise, all y values were again normalised to fit into the range [0, 1]. The number of bin divisions in both X and Y directions was varied from 5 to the values shown in Table 1. Each combination of function, points, bins and noise was repeated 20 times. The correlation measure was calculated as:

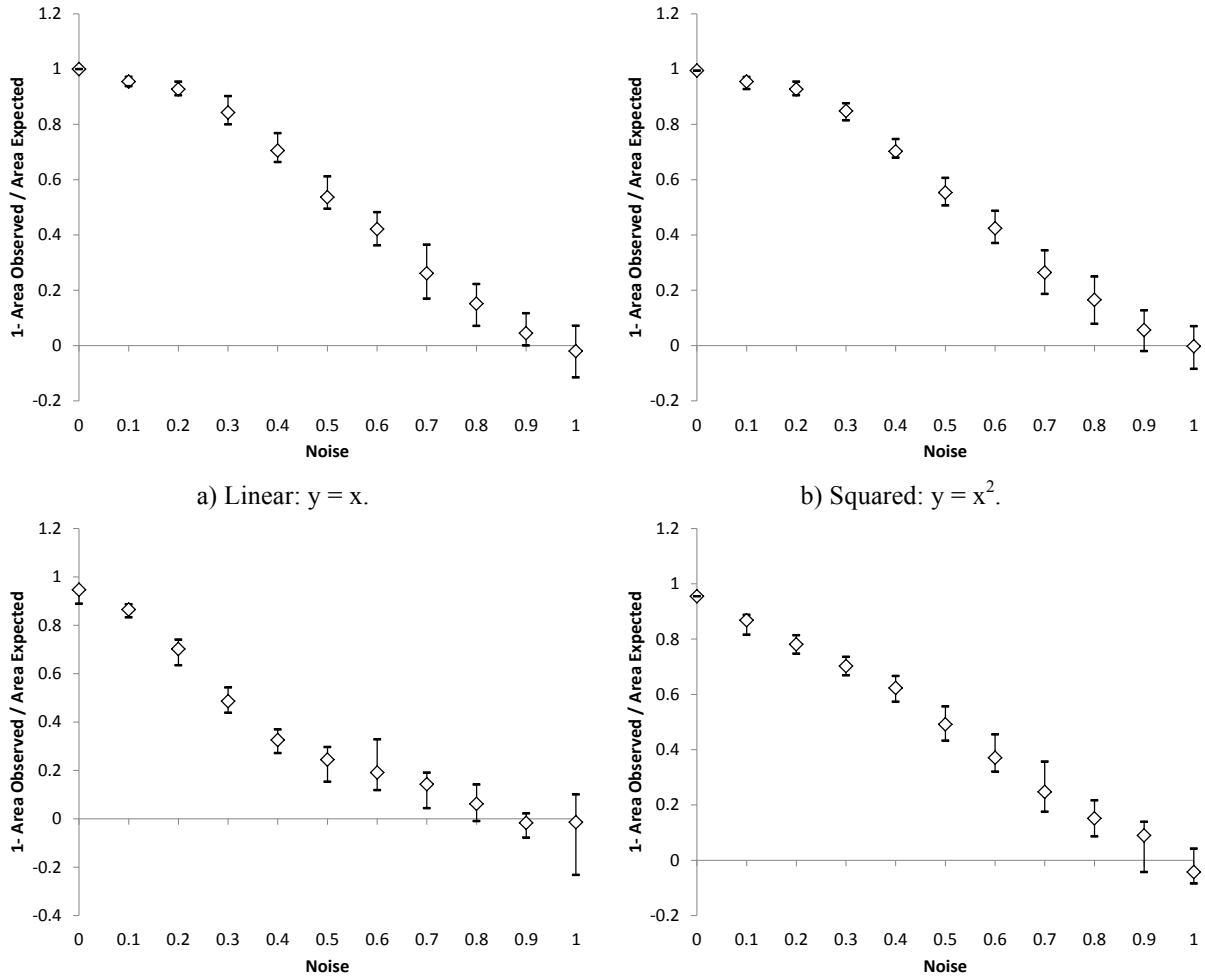
$$DARCC = 1 - \frac{\sum_{k=1}^{\lambda} \frac{C_k}{C_{tot}}}{\sum_{k=1}^{\lambda} \frac{\lambda^k e^{-\lambda}}{k!}}$$

An examination of results showed that the lowest error between DARCC and *v* occurred when the number of bins is set to the square root of the number of points, so that λ ≈ 1. All tests were repeated 20 times, and the median, 0.1 and 0.9 percentiles were calculated for each noise value.

In addition to these generated datasets, real data was selected from a dataset of the World Health Organisation (WHO), as reported in (Reshef et al., 2011). From Figure 4 of that paper, 4 datasets were selected where the number of points > 100. As DARCC is sensitive to outliers, points further than +-3 SDs from the mean in X or Y direction were removed. Results from DARCC were compared to those from MIC (Reshef et al., 2011).

## 6 Results

For the generated datasets using 500 points, out of all possible number of bins from 5 to 22, the results using 22 bins provided the lowest sum of squared error.




a) Linear: y = x.

b) Squared: $y = x^2$.

c) Cross: y = x or y = 1 - x

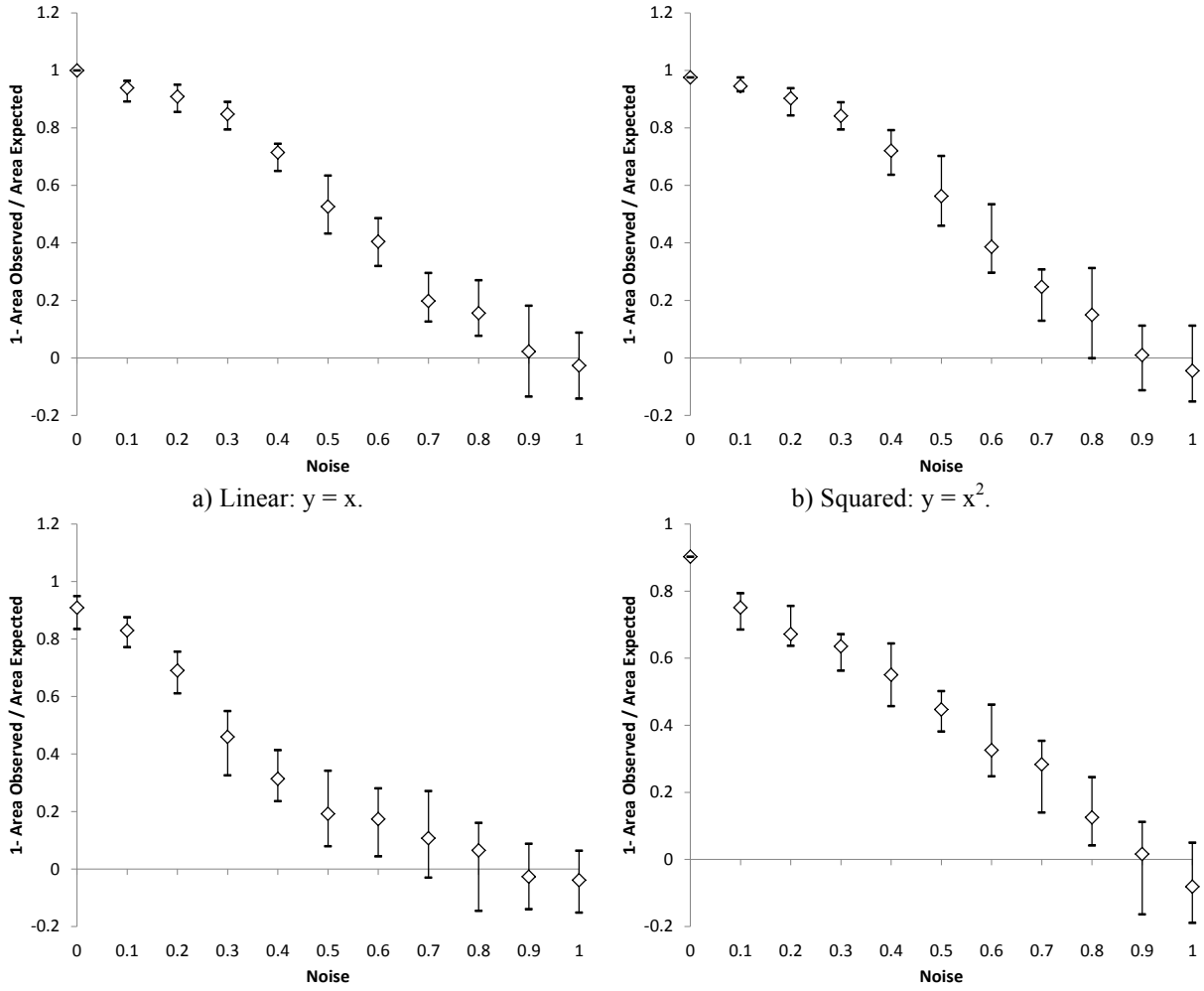
d) Sine: y = sin(2πfx), f = 2.7.

Figure 3. Relationship between proposed measure and noise for various functions, using 500 points and 22 bins. Graphs show 0.1, 0.5 and 0.9 percentiles.

Figure 3 shows results comparing the correlation measure with added noise magnitude, using 500 points. There is a good correlation between DARCC and the noise for all four functions examined. Results for the linear function (a) show that DARCC gives a higher value than expected when noise is above 0.5, meaning that a linear function will show a small bias toward a higher correlation coefficient. Likewise, the squared function (b) also shows a small bias towards higher correlation value. Conversely, the cross function (c) shows a bias towards reporting a lower correlation value than required. The sine function (d) shows the best result, with an almost linear relationship between added noise and correlation coefficient.

Figure 4 shows results from tests using 250 points. Out of all possible number of bins from 5 to 15, results using 15 bins provided the lowest sum of squared error. As with the results above, the

proposed measure is well correlated with the noise input for all four functions examined. Small deviations from the ideal relationship are similar to those for 500 points.




a) Linear: $y = x$.

b) Squared: $y = x^2$.

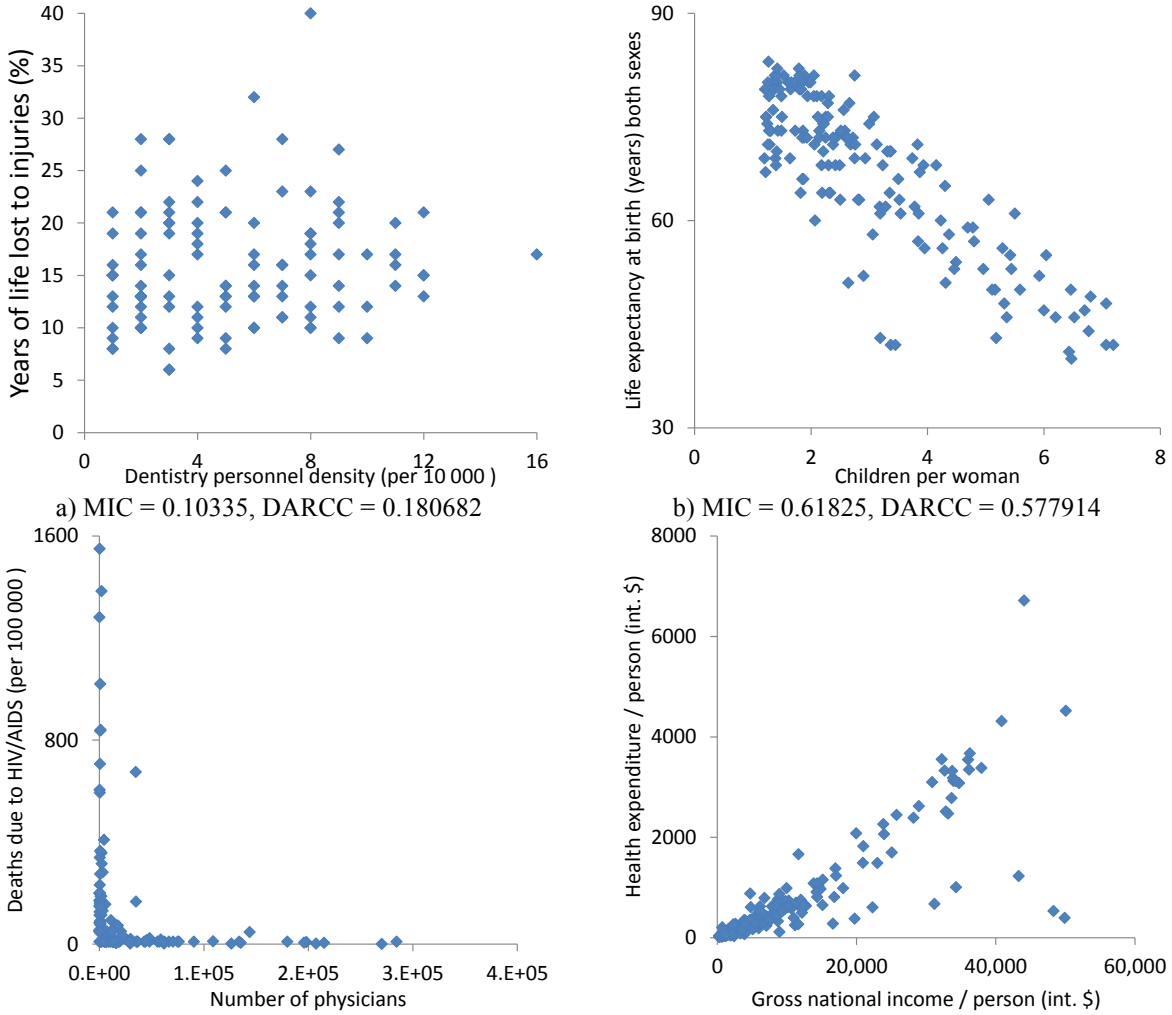
c) Cross: $y = x$ or $y = 1 - x$

d) Sine: $y = \sin(2\pi fx)$, $f = 2.7$.

Figure 4. Relationship between proposed measure and noise for various functions, using 250 points and 15 bins. Graphs show 0.1, 0.5 and 0.9 percentiles.

Comparing these results with the desirable properties listed in section 3, the first, that the measure must be bounded by [0,1] has been met. The second, that the measure should exhibit a linear response to added noise, is close to being achieved. The third, that the measure should be invariant over the type of function, is also close to being achieved. This is remarkable given the simplicity of the measure and the range of functions examined.

Results from real data sets are shown in Figure 5. For a) and b) there is good agreement between the MIC and the proposed measure. For c) and d) there is a wider difference. Some of this is due to the fact that MIC uses all the data, whereas DARCC removes outliers where x or y is +/- 3 standard deviations from the mean.




a) MIC = 0.10335, DARCC = 0.180682

b) MIC = 0.61825, DARCC = 0.577914

c) MIC = 0.6972, DARCC = 0.887355

d) MIC = 0.83774, DARCC = 0.576136

Figure 5. Some relationships based on data from the World Health Organization, after Reshef et al, 2011), showing the value of MIC and the proposed measure.

## 7 Conclusion

The proposed measure Distribution Area Ratio Correlation Coefficient (DARCC) is reasonably accurate and fast to calculate. It requires the input space to be partitioned into the same number of bins in both *x* and *y* directions, forming cells at the intersection of bins. The number of points falling into each cell is counted. The distribution of cells with *k* points is compared to the Poisson distribution. The optimum number of bin divisions was found to be equal to *floor(√p),* where *p* is the number of points in the data set. The area under the density curve is estimated for number of points per bins ranging from 1 to *λ*, where *λ* is the mean number of points per bin, or $p/bins^2$. The accuracy of the measure is not perfect, but this is balanced by its ease of computation. Preliminary results suggest this method may be orders of magnitude faster than MIC, for example. This measure may be suitable as a correlation coefficient for arbitrary functions, where a large number of variables must be compared in a limited time, or as a first approximation before other methods are employed. The question is not whether such simple methods may yield results of high accuracy, but whether some decrease in accuracy is justified by the high speed, when compared to other methods.

## 8 References

Breiman L and Friedman JH (1985) Estimating Optimal Transformations for Multiple Regression and Correlation, *J. Amer. Statist. Assoc*. 80, 580.

Darbellay G and Vajda I (1999) Estimation of the information by an adaptive partitioning of the observation space, *IEEE Trans. Inf. Theory* 45(4), 1315 – 1321.

Delicado P and Smrekar M (2009) Measuring non-linear dependence for two random variables distributed along a curve, *Statistics and Computing* 19, 255–269.

Hastie TJ and Stuetzle W (1989) Principal curves, *J Am Stat Assoc* 84, 502–516.

Kraskov A, Stögbauer H and Grassberger P (2004) Estimating mutual information, *Physical Review E Stat.Nonlin. Soft Matter Phys.* 69, 066138.

Moon YI, Rajagopalan B and Lall U (1995) Estimation of mutual information using kernel density estimators. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 52(3), 2318-2321.

Rényi A (1959) On Measures of Dependence, *Acta Math. Hung*. 10(3-4), 441.

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M and Sabeti PC (2011) Detecting Novel Associations in Large Data Sets, *Science* 334, 1518.

Székely G and Rizzo M (2009) Brownian distance covariance, *Annals of Applied Statistics* 3(4), 1236.

Tibshirani R (1992) Principal curves revisited, *Statistics and Computing* 2(4), 183-190.

Wang Q, Shen Y and Zhang JQ (2005) A nonlinear correlation measure for multivariable data set, *Physica D* 200, 287–295.