Quantifying Dependence

Ella Kaye Kaspar Märtens Paul Vanetti Andi Wang

University of Oxford
Department of Statistics
February 25, 2016

Abstract

Quantifying dependencies of random variables has been a much-discussed problem is recent years. In this report we review some proposed dependency measures; mutual information, maximal information coefficient and the distance correlation. We discuss the reproducibility of the analysis presented in Reshef *et al.*'s initial paper on the maximal information coefficient [5], and carry out some additional comparisons of statistical power. Finally, we discuss the algorithm proposed by Reshef *et al.* to calculate the maximum information coefficient and show that it often underestimates the true value.

1 Introduction

How can we measure the level of dependency between two random variables? This is an important question in data exploration, especially given the size of modern data sets and the enormous number of possible dependencies; far more than could be checked manually. To this end, we would like a statistic, which we can calculate easily from data, which returns 0 for independent data and larger values for data exhibiting dependency, regardless of how non-linear the dependency is.

2 Background Theory

In this section we will briefly review the methods of measuring dependence which we will be focusing on: mutual information (denoted I), maximal information coefficient (MIC) and distance correlation (dCor).

Let X, Y be two random variables, defined on the same probability space, with joint density $p_{X,Y}$ and marginal densities p_X and p_Y respectively. We want to quantify their level of dependence. The Pearson correlation R^2 is a well-known measure of linear dependence. R^2 can effectively describe dependence when two variables have an underlying linear relationship with homogeneous noise, but is unable to detect non-linear relationships. One of Reshef *et al.*'s aims [5] was to develop a statistic to measure dependence which exhibits *equitability*, giving "similar scores to equally noisy relationships of different types," not favouring any specific relationship such as linearity.

First we define the *mutual information (MI)* between $X, Y \sim p_{X,Y}$:

$$I[X;Y] = \mathbb{E}\left[\log_2 \frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right].$$

It can be shown that this is always non-negative, and equals 0 if and only if X and Y are independent. Hence $I[X;Y]$ will be strictly positive if there is any dependence between X and Y, regardless of what the dependency is. Intuitively, $I[X;Y]$ can be thought of as a measure of how much information one random variable contains about the other. It is closely related to other concepts from information theory such as the entropy[1], which is often referred to as the "self-information" of a random variable [2]. The mutual information is a property of the underlying distributions, and estimating it from a sample is non-trivial. Later on in our simulations we will use the k-nearest neighbours (KNN) method as in [4]. The choice of base 2 for the logarithm means that the units of I are bits[2]. A choice of base e (resulting in a multiplicative factor) would give units of nats.

In [5] the proposed statistic is the *maximal information coefficient* (MIC). Given a data sample $\{(x_i, y_i)\}_{i=1}^n$ where $x_i, y_i \in \mathbb{R}$ for each $i = 1, \ldots, n$, the MIC is calculated as follows. Up to a maximum data-dependent grid resolution, the largest possible mutual information achievable by any grid applied to the data is calculated. By this, we mean an estimate of the mutual information is calculated by looking at the empirical distribution of the grids; each grid has probability 'density' proportional to the number of points lying within it. These mutual information values are adjusted for different grid sizes and normalised to lie between 0 and 1. The authors provide choices for the maximum grid resolution.

From an applied perspective, some weaknesses of the MIC statistic are that there is some tuning required; the choice of maximum grid resolution $B(n)$, and the fact that it doesn't scale naturally to multidimensional data. The algorithm by Reshef *et al.* involves sorting the data, which is not possible for non-scalar data.

Another dependence measure from [7] is the *distance correlation (dCor)*. Let $f_{X,Y}, f_X, f_Y$ denote the joint and marginal characteristic functions of (X, Y), X and Y respectively. The distance correlation is based on the fact that if X, Y are independent, then the joint characteristic function factors $f_{X,Y} = f_X f_Y$. Then a natural distance to consider is $\|f_{X,Y} - f_X f_Y\|$ where $\|\cdot\|$ is some suitable norm on the function space. The authors give a specific choice of norm which leads to many pleasing properties of the statistic, and a straightforward method to estimating the distance correlation from a sample.

3 Experiments and results

In [5], Reshef *et al.* apply their algorithm for calculating MIC to four real-world datasets, and compare its performance to other measures of dependence. In this section, we first discuss our attempts to replicate their results and then carry out some additional comparisons. Finally, we discuss the approximate algorithm for calculating MIC and demonstrate that often this scheme underestimates the true MIC value.

Our code for figures and analysis is available in `https://github.com/pjcv/mic`.

[1] $H(X) = \mathbb{E}[-\log_2(p_X(X))]$

[2] So, for instance, the entropy of a fair coin is 1 bit.

3.1 Replicating Reshef et al (WHO data set)

Reshef *et al.* [5] consider a data set of 357 social, economic, health and political indicators from the World Health Organisation, for 202 countries. The data as used in the paper are available from www.exploredata.net, a website about MINE maintained by Reshef & Reshef. Here we describe our attempts to replicate their Figures 4A and 4B from [5], and raise some concerns about some other elements of the data visualisation choices in Figures 4C-4I. The original plot is shown in Figure 10 in Appendix A (we will continue to refer to it as Figure 4, to match its original name).

Of the 72114 possible data values, 22856 are missing (32%). At no point in the paper or the supplementary materials do the authors mention this missing data or describe their strategy for dealing with it. Experimentation with their function revealed a simple approach: each pair of variables are only compared on the samples they have in common. When processing this data set, they required that at least 25% of the samples be shared (i.e. that both of the pair of variables had been measured on a minimum of the same 50 countries). They offer no justification as to why they chose 25%. One criticism of [5] is that the default choice of $B(n) = n^{0.6}$ as the choice of setting the grid resolution in the MIC algorithm is arbitrary/unjustified. Here, that default is changed to $B(n) = n^{0.65}$, but again, no justification for this figure, or for changing the default, is given. In what strikes us a violation of good data visualisation principles, for Figure 4I, the algorithm was run with $B(n) = n^{0.7}$, so that on the same data set, within the same figure, different variants of the algorithm were used, with no explanation.

To replicate Figure 4A, we ran the complete WHO dataset through the MINE software available at www.exploredata.net, with the changes to the default parameters, as described above. We note that there is no label on the legends in Figures 4A and 4B. Given that the figures are comprised of tens of thousands of (overlapping) points, and given the scale, we take it that the colour represents the log of the counts of number of points in that area of the graph.

Our plot of MIC vs. the Pearson Correlation coefficient, produced from the output[3] of running the Reshefs' software, with the parameters used in their paper, is shown in Figure 1 *(left)*. We experimented with several different bin sizes, but, using the `ggplot2` library, this is the closest, visually, we could get to Reshef *et al.* Figure 4A. Note that there appears to be more scatter in this plot than the original. Our tick-marks on the colour-bar are given as the log of the counts, and corresponds roughly to the colours and intensities in the original plot. To see if we could replicate the plot without the additional scatter, we ran the software again, looping over each pair of variables, this time keeping the default arguments in the function call, but rejecting within the loop any pair where there were fewer than 25% of the samples that had data for both. The result is shown in Figure 1 *(right)*. We would have expected the two graphs in Figure 1 to look the same, but in fact the plot on the right appears much closer to the original Figure 4A.

Figure 4B represents further challenges to reproducibility. The mutual information has been calculated through the algorithm of Kraskov *et al.* [4]. This is a k-nearest neighbours algorithm, but in neither the paper ([5]) nor its supplementary information do Reshef *et al.* give the value of k used in generating the data for this figure. Nor do the authors state which unit of mutual information they have used.

We estimated the mutual information using both bits and nats, with $k = 1, 6, 10$, using the `R` package `FNN`. The closest of these to replicating Figure 4B is $k = 1$, measured in nats. The result is shown in Figure 2. Although there are over 60,000 possible pairs, this graph represents far

[3] Confusingly, in the output generated by the software, the column that contains the Pearson correlation coefficient is headed "linear regression (p)".

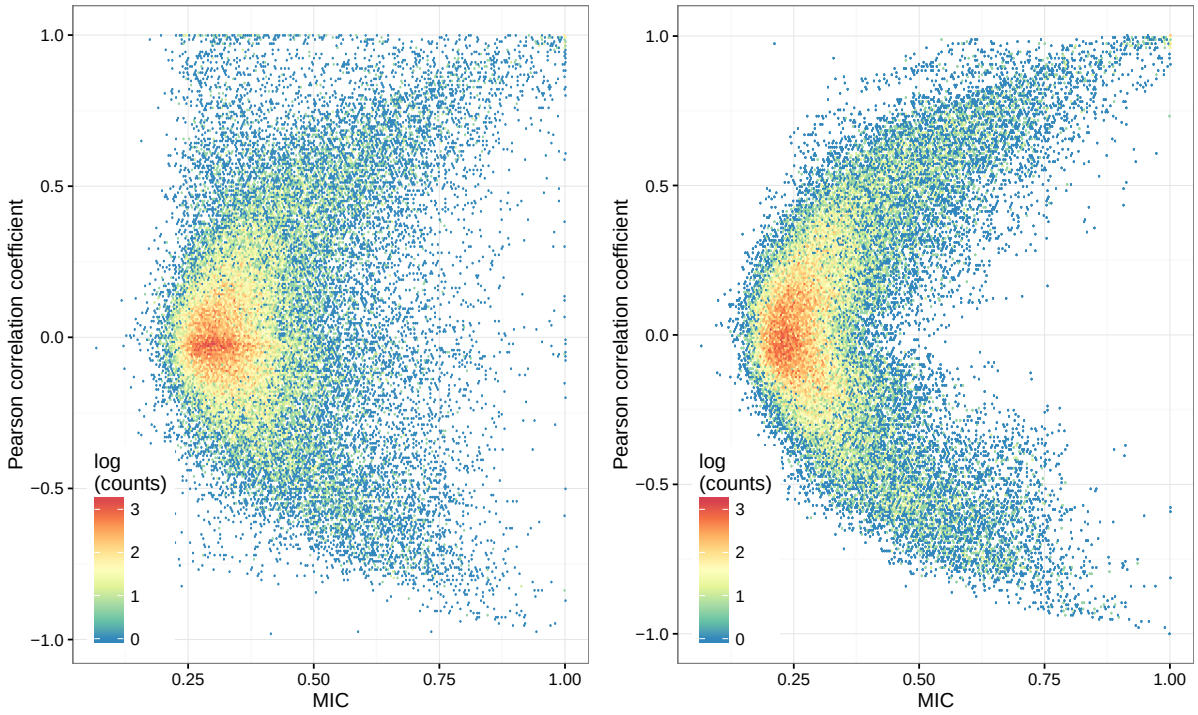

Figure 1: Attempt to replicate Figure 4A, MIC against the Pearson correlation coefficient, from [5]. *(left)* Using same function settings as the paper. *(right)* The same plot, after adjustment, looks more like the one that was published.

fewer points. 13343 pairs were lost through having less than 50 samples in common. A further 14730 produced 'NaN's for the mutual information estimate, and were excluded from the graph. A clear concern with this plot is that there are many (6834) points where the mutual information was estimated as less than zero, when by definition it must be non-negative. In Figure 4B from [5], the x-axis starts at 0, but it is no clear whether the authors have discarded the negative results, set them equal to zero, or something else entirely.

One further concern with Figure 4 in [5] is with sub-figures 4F and 4H. In each of these figures, two trends are identified. They were found by considering the non-linearity measure $MIC - \rho^2$, which in this case is 0.5011, only just lower than its MIC score of 0.5029. For 4F, which plots income per person against adult (female) obesity, the points are split into a linear trend, described in the figure caption as comprising "a set of Pacific Island nations in which obesity is culturally valued" whilst "most other countries follow a parabolic trend." However, Table S10 in the supplementary material reveals they have included Egypt and Iraq within the former trend. So, the first trend is of eight countries (of which 25% do not correspond to the phenomenon they claim the trend captures), whilst 85 countries contribute to the main trend. However a cursory glance of Figure 4F certainly makes it appear that the importance/weight of the two trends is more equally balanced than that. This plot could easily have been explained as one trend with a group of four outliers.

3.2 Comparisons on yeast gene expression data

Here we explore the behaviour of various dependency measures on a yeast gene expression data set used in [5], where MIC was compared to a special-purpose statistic developed by Spellman *et al* [6]. The data consists of 4381 time series, where gene expression has been measured at 23

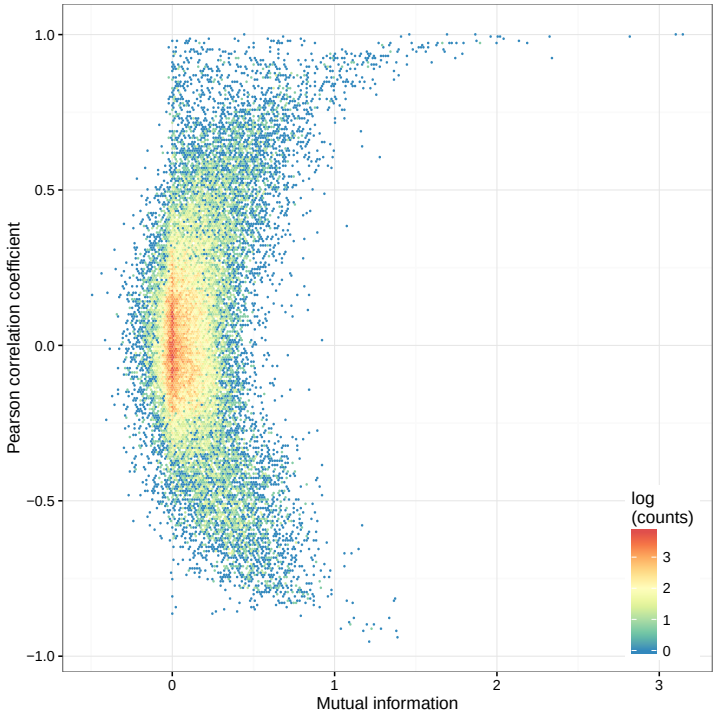

Figure 2: Mutual information (in nats) against the Pearson correlation coefficient. Mutual information estimated using KNN [4] with $k = 1$. Attempt to replicate Figure 4B from [5].

time points, and it is of interest to detect which genes exhibit changes in transcription levels over time. We have chosen to compare MIC against the methods introduced in Section 2 (R^2; mutual information estimated via KNN, denoted by MI(k); and dCor) using this dataset; this is opposed to the artificial dependencies such as circular or checkerboard patterns used in [3]. For the k-nearest neighbors based algorithm we used $k = 1$ and $k = 5$ [4].

The strongest associations by each method are shown in Figure 3 together with their respective ranks (out of 4381) as computed by the other methods. In general, most associations shown there are ranked relatively high by every method, with an exception of the fourth example for MI($k = 1$). We also note that MIC estimation results in ties, all the top four associations having an equal MIC value.

Figure 4 compares the dependency scores of MIC and other methods for all genes. They are positively correlated, but different approaches provide clearly different orderings for the strength of association. We also note that even though the mutual information is non-negative by definition, we have obtained negative estimates for this (Fig. 4 panels for MI($k = 1$) and MI($k = 5$)).

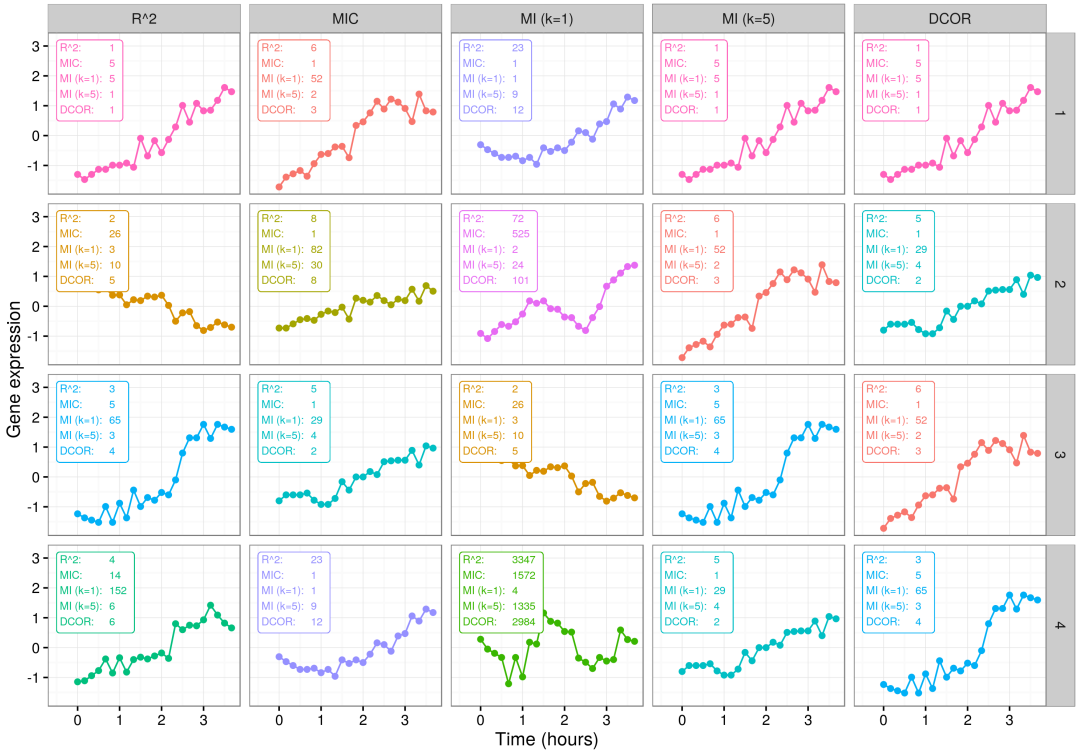

Figure 3: Comparison of methods for detecting associations in the yeast gene expression data set. Top 4 hits based on the relative ranking of various methods (Pearson squared correlation R^2, MIC, mutual information estimated via KNN ($k = 1, k = 5$), and Brownian distance correlation). The same associations (i.e. genes) are shown in the same colour, and the legends indicate the relative ranks assigned by the various methods.

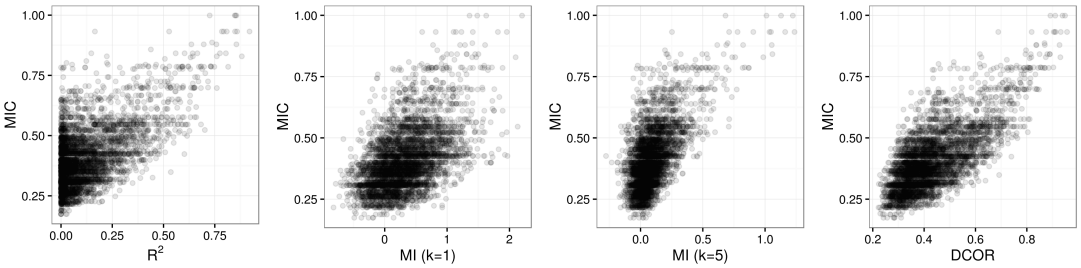

Figure 4: Scatter plots to compare MIC (y-axis) against various methods (x-axis). Each point denotes the respective dependency score for one gene.

3.3 Comparing Statistical Power

In Fig. 4 of [3] the authors compare the statistical power of numerous statistics for various types of relationships and for different noise levels. They considered linear, parabolic, sinusoidal, circular, and checkerboard relationships. We were able to reproduce the linear plot, and wanted to make a similar comparison for a new type of relationship. We chose to use a Raichu-like relationship [1], see Fig. 5. We compared the power of the following statistics: R^2, dCor, MI for $k = 1, 6, 20$ and MIC. We used the R packages `energy` (dCor), `FNN` (k-nearest neighbours for MI) and `minerva` (MIC).

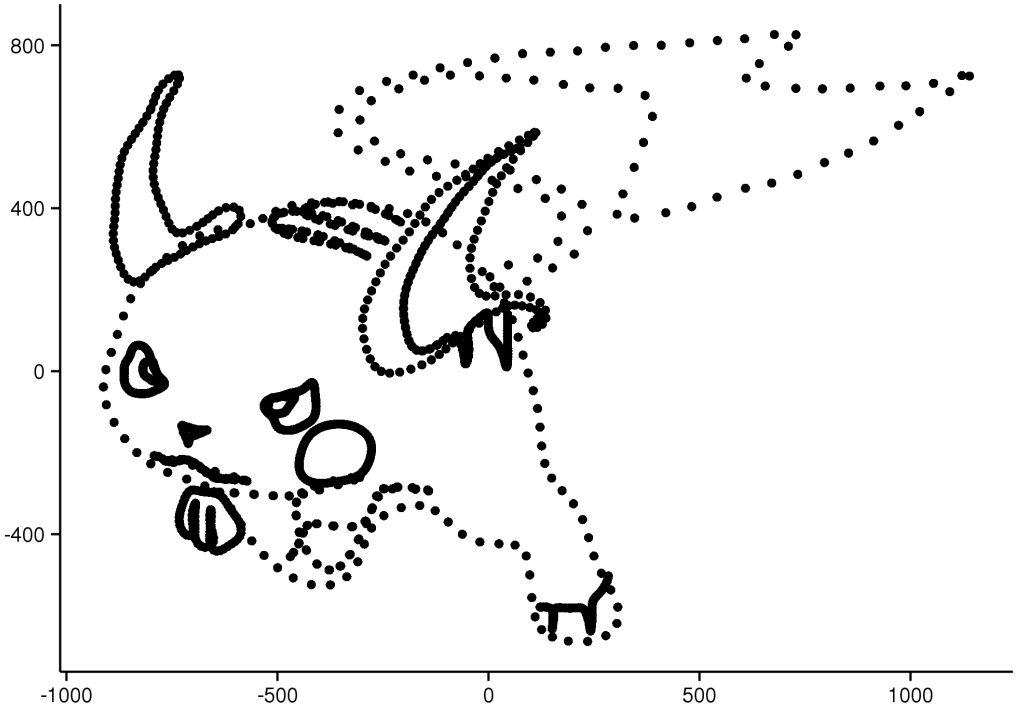

Figure 5: 3000 points of the Raichu-like curve, with no noise added.

The Cartesian coordinates of the underlying Raichu-like curve from [1] are given parametrically, as (necessarily complex) functions $x_{\mathrm{R}}(t), y_{\mathrm{R}}(t)$ of a parameter t which ranges from 0 to 68π. To produce Figure 2 we took 3000 points $t_1, \ldots, t_{3000}$ uniformly spaced in $[0, 68\pi]$ and plotted the corresponding $\{(x_{\mathrm{R}}(t_i), y_{\mathrm{R}}(t_i))\}_{i=1}^{3000}$.

By "statistical power" we mean the probability that a statistic, when evaluated on data exhibiting a true dependence, produces a value which is significantly different to that obtained when applying the statistic to independent data. To test the power of various statistics applied to the Raichu-like relationship, we did the following, as in [3]. We took 2000 uniformly spaced points in $[0, 68\pi]$, and calculated the corresponding points of the curve. We removed erroneous values (for certain t values $x_{\mathrm{R}}(t)$ and $y_{\mathrm{R}}(t)$ are not real-valued), and then added Gaussian noise of a given amplitude to the y_{R}-values. We then applied the statistics to this noisy data. For comparison, we randomly permuted the noisy y_{R}-values to produce "null" data, to which we also applied the statistics. We repeated this 100 times independently. The power is estimated by the proportion of the statistics above the threshold determined by the 95th percentile of the statistics obtained from the null data.

The results are shown in Figure 6. All statistics appear to exhibit reasonable power, even R^2, despite the large amplitudes of noise added. In this case, MIC and dCor seem to have slightly higher power compared to alternatives.

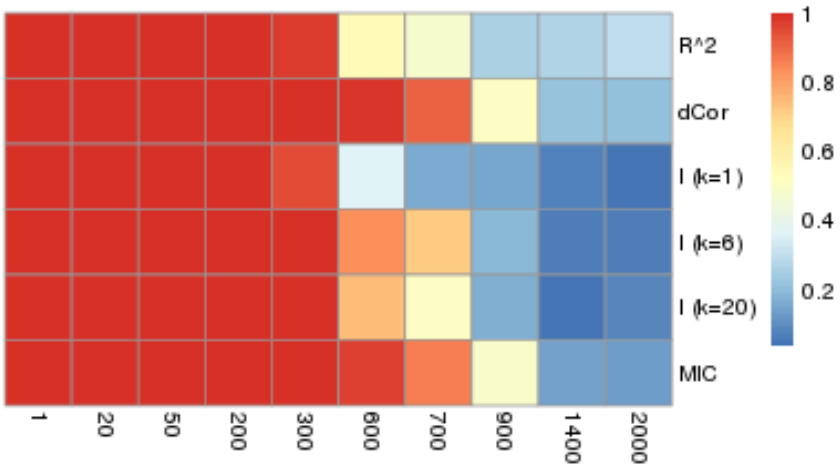

Figure 6: Heatmap of the statistical power (indicated by color) of various methods (y-axis) for the Raichu-like curve, for different noise levels (the standard deviation of the Gaussian noise added to the y_{R}-components on the x-axis). Here, mutual information is denoted as I.

3.4 Heuristic Approximation

Computing MIC is not straightforward. Searching for the maximum mutual information over all possible partitioning arrangements is computationally intensive - for a given number of columns x and rows y, the number of possible grids is $\binom{N}{x-1} \times \binom{N}{y-1}$. While the authors of [5] do not state precise results regarding the computational complexity of their method, they provide (and presumably employ) a heuristic algorithm which only approximates their definition of MIC.

[5] describes a computationally efficient method for choosing the partition in *one dimension only* using dynamic programming techniques. For a given number of columns and rows, they first "equipartition" one dimension; for that dimension, all rows (or columns) will contain the same number of data points. (If the data points are not evenly divisible, some partitions contain one more data point.) Given this equipartitioning they then find the optimal partitioning in the other dimension using their dynamic-programming scheme. They consider the maximum over equipartitioning in both dimensions to ensure symmetry.

This scheme clearly underestimates the "true" MIC, as the maximum mutual information is taken over only a subset of grids. In order to determine the magnitude of this discrepancy, we implemented the "exact" MIC algorithm by exhaustively searching over all possible grids. We refer to this algorithm as *MIC-exact*, and the algorithm of [5] as *MIC-heuristic*. We apply both algorithms to the yeast data set as above [6]. For each gene we have $N = 23$, so the bound on partitions $B(23) = 6.56$ meaning we only need to test $(x, y) \in \{(2, 2), (2, 3), (3, 2)\}$ which is feasible to compute with MIC-exact. The comparison between exact and heuristic MIC values is shown in Figure 7.

These results suggest a significant discrepancy due to the approximation made in the heuristic algorithm. To illustrate this further, we show in Figure 8 the gene with the largest difference between the exact and heuristic implementations, and in Figure 9 the gene with the largest difference and with an exact score of greater than 0.9. The latter case suggests a clear relationship which is not captured by the equipartitioning scheme.

The discrepancy between MIC-exact and MIC-heuristic raises some questions about the properties of MIC:

- **Equitability.** Does the heuristic approximation diminish the equitability of MIC? This seems likely, as those relationships not represented in an equipartitioning model may not be captured accurately.
- **Power.** The heuristic approximation generally reduces MIC scores. Does this serve to increase the power of the statistic?

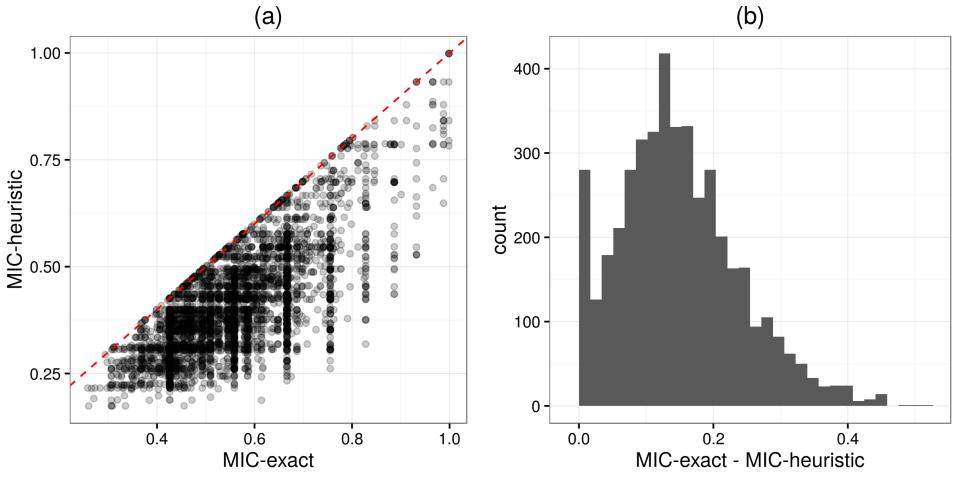

Figure 7: (a) MIC-exact (x-axis) vs MIC-heuristic (y-axis), with each point representing one gene of the Spellman yeast dataset [6]. The red dashed line shows identity $y = x$. (b) The distribution of differences between the exact MIC and its approximation for the same data.

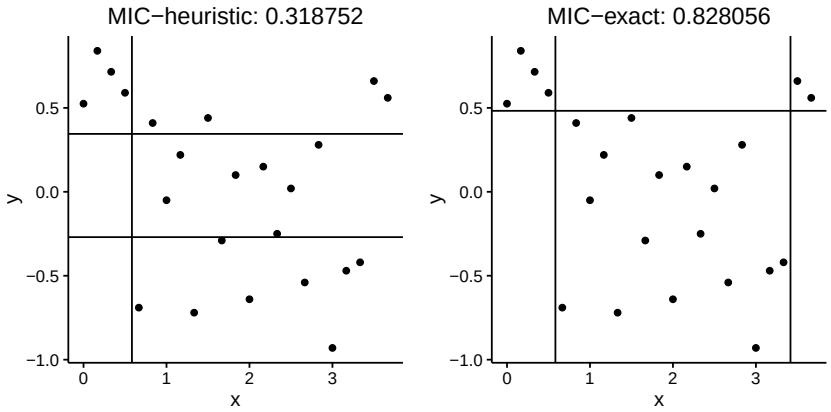

Figure 8: Partitions found by the heuristic and exact algorithms for the gene with the largest discrepancy due to the heuristic approximation.

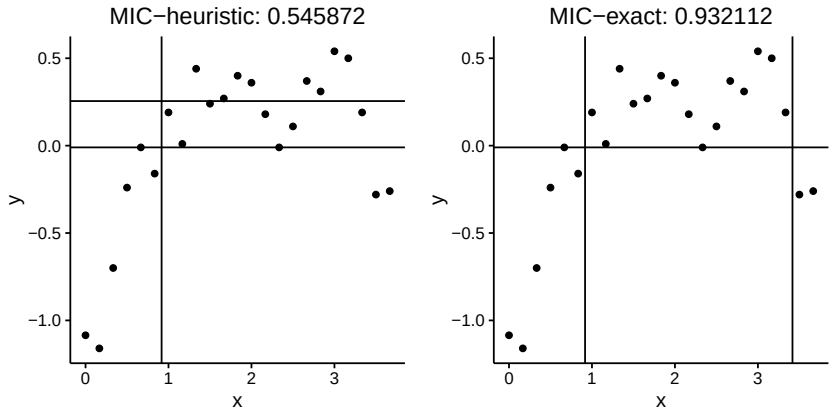

Figure 9: Partitions found by the heuristic and exact algorithms for the gene with the largest discrepancy due to the heuristic approximation, where the exact MIC is found to be > 0.9.

References

[1] Wolfram Alpha. Raichu-like curve. `https://www.wolframalpha.com/input/?i=raichu-like+curve`. [Online; accessed 24-Feb-2016].

[2] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, New York.

[3] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[4] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[5] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[6] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[7] Gábor J Székely, Maria L Rizzo, et al. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

A Plots from papers under discussion

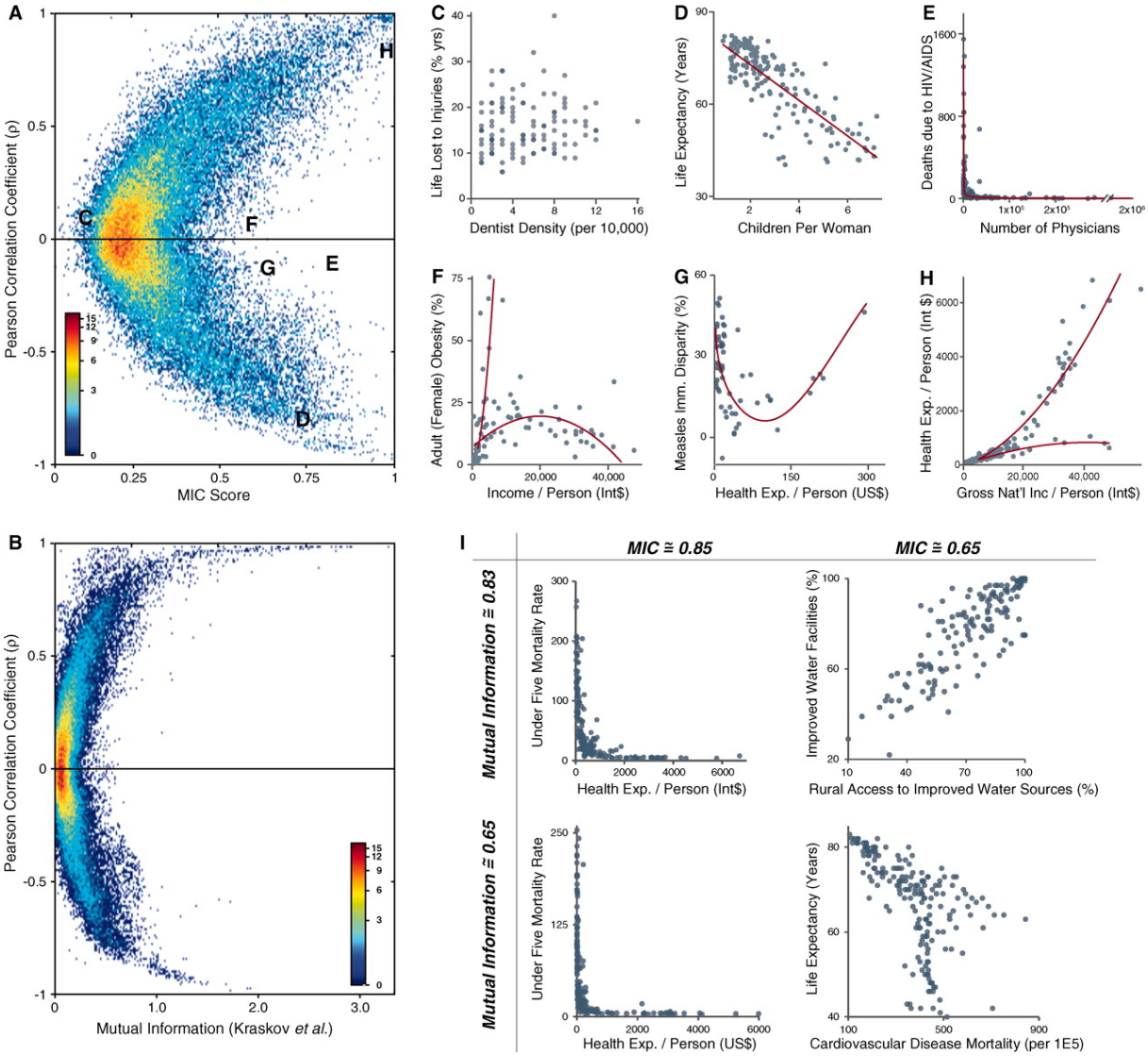

Figure 10: Figure 4 from Reshef *et al.* [5]