Detecting Novel Associations in Large Data Sets

(Reshef et al, *Science* 334: 1518-1524, 2011 + Kinney and Atwal, 2013)

Dino S.

Gatsby Unit

February 14, 2013

Detecting nonlinear associations

- Aim: identify "interesting relationships" between pairs of (scalar) variables in "large data sets"

Detecting nonlinear associations

- Aim: identify "interesting relationships" between pairs of (scalar) variables in "large data sets"
- functional relationships and their "superpositions"

Detecting nonlinear associations

- Aim: identify "interesting relationships" between pairs of (scalar) variables in "large data sets"
- functional relationships and their "superpositions"
- **equitability**: "similar scores to equally noisy relationships of different types"

Detecting nonlinear associations

- Aim: identify "interesting relationships" between pairs of (scalar) variables in "large data sets"
- functional relationships and their "superpositions"
- **equitability**: "similar scores to equally noisy relationships of different types"
- for functional relationships: roughly R^2 relative to regression function

Detecting nonlinear associations

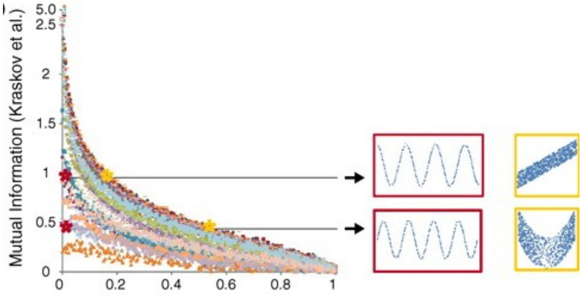
- Aim: identify "interesting relationships" between pairs of (scalar) variables in "large data sets"
- functional relationships and their "superpositions"
- **equitability**: "similar scores to equally noisy relationships of different types"
- for functional relationships: roughly R^2 relative to regression function
- Maximal Information Coefficient (MIC)

You kind of mean dependence, right?

- $I(X; Y) = 0$ iff $X \perp\!\!\!\perp Y$

You kind of mean dependence, right?

- $I(X;Y) = 0$ iff $X \perp\!\!\!\perp Y$
- Reshef et al: standard estimators (k-NN, Kraskov, Stogbauer and Grassgerger, 2004) of MI *do not* satisfy equitability



sample size: 500

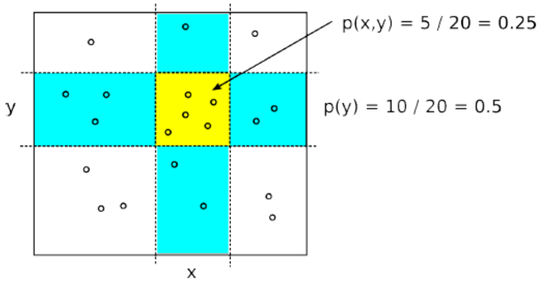
Binning the data

- *"...a grid can be drawn on the scatterplot that partitions the data to encapsulate that relationship..."*

Binning the data

- *"...a grid can be drawn on the scatterplot that partitions the data to encapsulate that relationship..."*

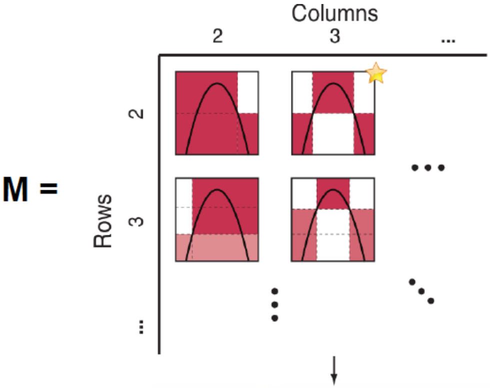
n = 20



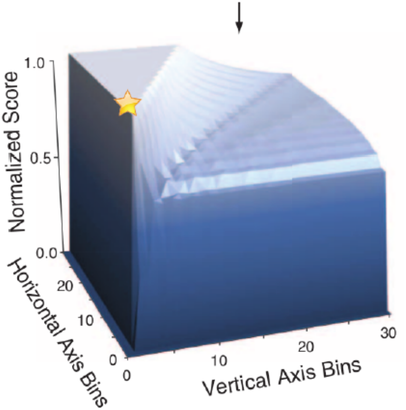
- naïve MI estimate: $I(X;Y) \approx I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
- $I(x;y) \leq \min\{H(x), H(y)\} \leq \log_2(\min\{n_x, n_y\})$

Binning the data (2)

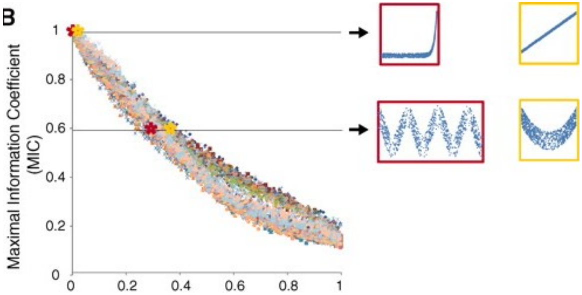
- explore grids of n_x-by-n_y resolution $\rightarrow$ find a grid that maximizes mutual information & normalize $\rightarrow$ output a number m_{n_x,n_y}



Binning the data (3)



$$MIC[X;Y] = \max_{n_x n_y \leq B} m_{n_x,n_y} = \max_{n_x n_y \leq B} \max_{n_x \times n_y \text{ grids}} \frac{I(x;y)}{\log_2(\min\{n_x, n_y\})}$$



sample size: 500

MATHEMATICS

A Correlation for the 21st Century

A novel statistical approach has been developed that can uncover nonlinear associations in large data sets.

Terry Speed

Most scientists will be familiar with the use of Pearson's correlation coefficient r to measure the strength of association between a pair of variables: for example, between the height of a child and the average height of their parents ($r \approx 0.5$; see the figure, panel A), or between wheat yield and annual rainfall ($r \approx 0.75$, panel B). However, Pearson's r captures only linear association, and its usefulness is greatly reduced when associations are nonlinear. What has long been needed is a measure that quantifies associations between variables generally, one that reduces to Pearson's in the linear case, but that behaves as we'd like in the nonlinear case. On page 1518 of this issue, Reshef *et al*. (*1*) introduce the maximal information coefficient, or MIC, that can be used to determine nonlinear correlations in data sets equitably.

Ysidro Edgeworth and later Karl Pearson gave us the modern formula for estimating r, and it very definitely required a manual or electromechanical calculator to convert 1000 pairs of values into a correlation coefficient. In marked contrast, the MIC requires a modern digital computer for its calculation; there is no simple formula, and no-one could compute it on any calculator. This is another instance of computer-intensive methods in statistics (*3*).

It is impossible to discuss measures of association without referring to the concept of independence. Events or measurements are termed probabilistically independent if information about some does not change the probabilities of the others. The outcomes of successive tosses of a coin are independent events: Knowledge of the outcomes of some tosses does not affect the probabilities for the outcomes of other tosses. By convention, any measure of association between two variables must be zero if the variables are independent. Such measures are also called measures of dependence. There are several other natural requirements of a good measure of dependence, including symmetry (*4*), and statisticians have struggled with the challenge of defining suitable measures since Galton introduced the correlation coefficient. Many novel measures of association have been invented, including rank correlation (*5*, *6*); maximal linear correlation after transforming both variables (*7*), which has been rediscovered many times since; the curve-based methods reviewed in (*8*); and, most recently, distance correlation (*9*).

To understand where the MIC comes from, we need to go back to Claude Shannature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2013 > February > Article

NATURE | NEWS

Tangled relationships unpicked

A statistical method discovers hidden correlations in complex data.

Philip Ball

15 December 2011

The US humorist Evan Esar once called statistics the science of producing unreliable facts from reliable figures. An innovative technique now promises to make those facts a whole lot more dependable.

print
email
rights & permissions
share/bookmark

Brothers David Reshef of the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, Yakir Reshef, now at the Weizmann Institute of Science in Rehovot, Israel, and their coworkers have devised a method to extract from complex sets of data relationships and trends that are invisible to other types of statistical analysis. They describe their approach in *Science* today[1].

"This appears to be an outstanding achievement," says Douglas Simpson, a statistician at the University of Illinois at Urbana–Champaign. "It opens up whole new avenues of inquiry."

Dizzying complexity

Here is the basic problem. You have collected lots

Journal home
Current issue
For authors

E-alert

Number of
published
… and mor

Recent

1. Eco-engine
Nature | 13
2. Theatre: Th
Nature | 13
3. Policy: Clas
Nature | 13
4. Seven days
Nature | 13
5. Dual collisi
Nature | 13

natur

"a method to extract from complex sets of data relationships and trends that are invisible to other types of statistical analysis"

Equitable? (Kinney and Atwal 2013)

- No explicit definition given by Reshef et al

Equitable? (Kinney and Atwal 2013)

- No explicit definition given by Reshef et al

Model: $Y = f(X) + \eta$, where η can depend on X through $f(X)$ only, i.e., $X \to f(X) \to Y$ forms a Markov chain.

Definition (*Reshef et al* notion of R^2-equitability)

In the large data limit: $D[X;Y] = g\left(R^2\left[f(X);Y\right]\right)$ for some function g that does not depend on f.

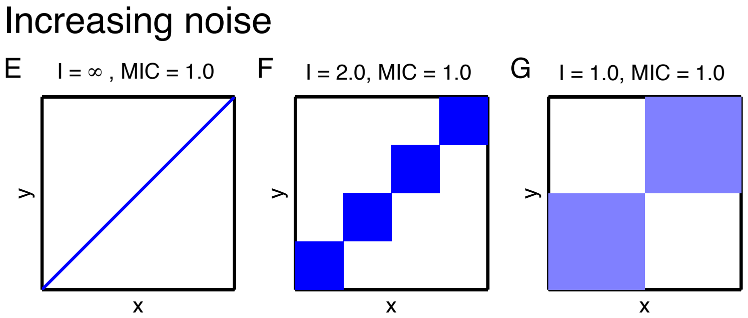
Not very equitable

- Kinney and Atwal (2013) prove: R^2-equitability is impossible for non-trivial D.

Not very equitable

- Kinney and Atwal (2013) prove: R^2-equitability is impossible for non-trivial D.
- **Proof**: Let $Y = X + \eta$, with $\eta \perp\!\!\!\perp X$, and let h be any invertible function. We can then write $Y = h(X) + \eta'$, where $\eta' = h^{-1}(h(X)) - h(X) + \eta$ is a valid noise term, as it depends on X through $h(X)$ only. Therefore, one should have $g\left(R^2\left[X; Y\right]\right) = g\left(R^2\left[h(X); Y\right]\right)\ \forall h$, which is impossible as R^2 is not invariant to general invertible transformations. Thus g, and therefore D, do not depend on the data!

Not very equitable (2)



Alternative notion by Kinney and Atwal (2013)

- **Self-equitability** (SE): $D[X; Y] = D[f(X); Y]$ whenever $X \to f(X) \to Y$ (log-pints vs. squared tea spoons)
- **Data-processing equitability** (DPE): For a Markov chain $X \to Z \to Y$, $D[X; Y] \leq D[Z; Y]$, i.e., processing cannot increase dependence.
- DPE $\implies$ SE

Alternative notion by Kinney and Atwal (2013)

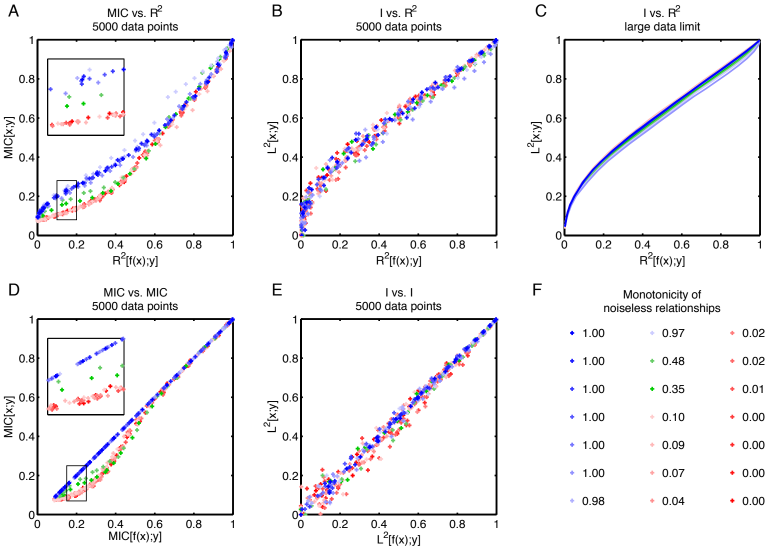
- **Self-equitability** (SE): $D[X;Y] = D[f(X);Y]$ whenever $X \to f(X) \to Y$ (log-pints vs. squared tea spoons)
- **Data-processing equitability** (DPE): For a Markov chain $X \to Z \to Y$, $D[X;Y] \leq D[Z;Y]$, i.e., processing cannot increase dependence.
- DPE $\implies$ SE
- MIC violates both, MI satisfies both

Alternative notion by Kinney and Atwal (2013)

- **Self-equitability** (SE): $D[X;Y] = D[f(X);Y]$ whenever $X \to f(X) \to Y$ (log-pints vs. squared tea spoons)
- **Data-processing equitability** (DPE): For a Markov chain $X \to Z \to Y$, $D[X;Y] \leq D[Z;Y]$, i.e., processing cannot increase dependence.
- DPE $\implies$ SE

- MIC violates both, MI satisfies both

- *We discuss the notion of equitability as a desirable heuristic property, as underscored by our use of words like "roughly equal" and "similar" instead of "equal" when discussing it. Philosophically, we have been using equitability as an approximate property.* (M. Mitzenmacher on A. Gelman's blog)



- $L^2[X;Y] = 1 - 2^{-2I[X;Y]}$ (Kinney and Atwal, 2013): *"...the simulation evidence offered by Reshef et al was artifactual..."*

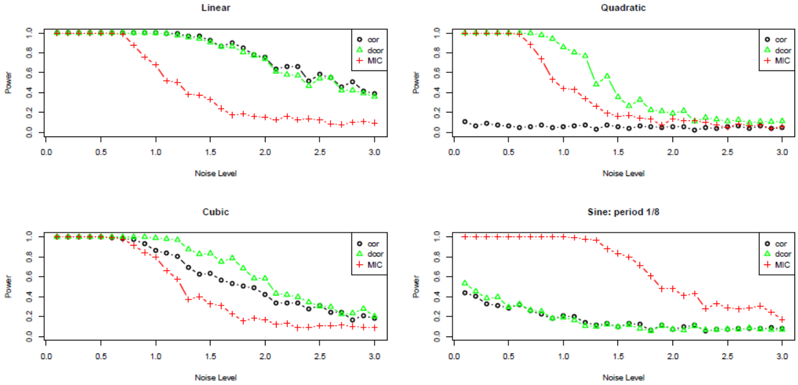
So what if it is not *really* equitable?

- It's a "powerful" technique anyway...

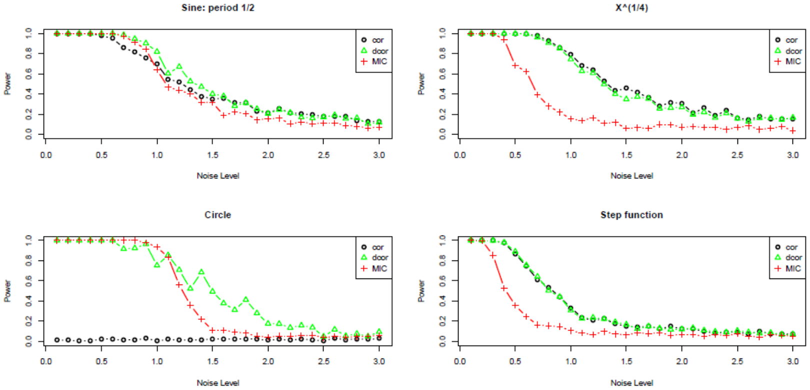
So what if it is not *really* equitable?

- It's a "powerful" technique anyway...
- ...at least philosophically.

MIC vs. dCor



- Simon and Tibshirani (2011): *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****.*



- Simon and Tibshirani (2011): *MIC has* ***serious power deficiencies****, and hence when it is used for large-scale exploratory analysis it will produce* ***too many false positives.***

At least it's fast to compute?

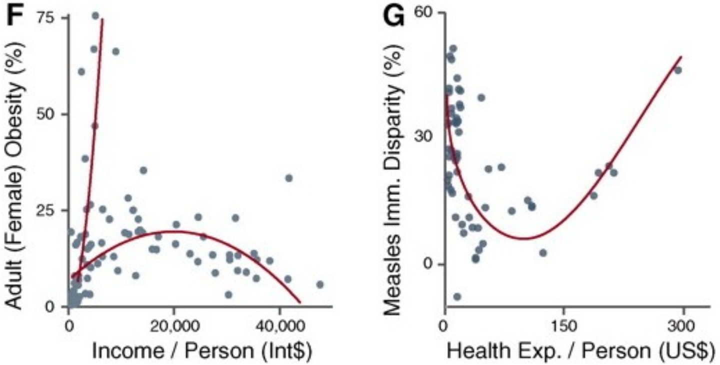
- Science magazine podcast: *It really can be applied to **just about any** data set.* (D. Reshef)

At least it's fast to compute?

- Science magazine podcast: *It really can be applied to **just about any** data set.* (D. Reshef)
- **Large** data sets: many (x-y) pairs of variables with $D = 1$, $N \leq 1000$.

At least it's fast to compute?

- Science magazine podcast: *It really can be applied to **just about any** data set.* (D. Reshef)
- **Large** data sets: many (x-y) pairs of variables with $D = 1$, $N \leq 1000$.
- *We observed the MIC algorithm of Reshef et al. to run* ~ 600 *times slower than the Kraskov er al. mutual information estimation...* (Kinney and Atwal 2013)


F
75
50
25
0
Adult (Female) Obesity (%)
0
20,000
40,000
Income / Person (Int$)
G
60
30
0
Measles Imm. Disparity (%)
0
150
300
Health Exp. / Person (US$)