Discussion of the Maximal Information Coefficient

Alexander W Blocker
http://www.awblocker.com/

Feb 21 2012

Outline

Outline

Motivation

- Have high-dimensional dataset
- 100s-1000s of variables; often fewer observations than variables
- **Goal:** find novel bivariate relationships
- General definition of relationships (not just nonlinear, even nonfunctional)
- "Equitable" wrt different types of relationships
- Alternative to manual search (according to authors)

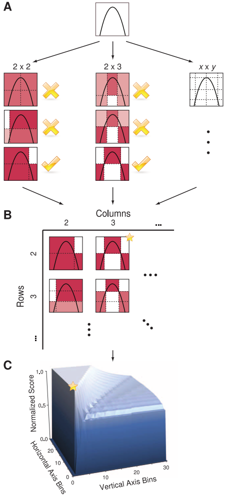
Generality & equitability

Stated goals of the method (heuristic)

- *Generality:* ability to detect broad range of relationships
  - Includes nonfunctional
  - Also want "noncoexistence" and mixtures of functions
- *Equitability:* similar scoring of "equally noisy relationships of different types"
  - Harder to pin down; asymptotic?
  - How do nonfunctional fit?
  - Symmetry $\rightarrow$ complications; predictive distribution from sinusoid, e.g.

Technical definition

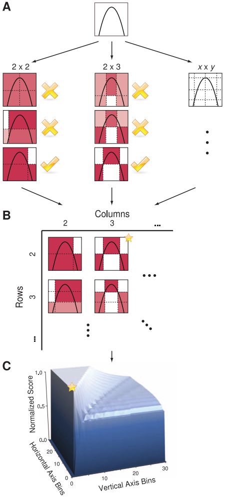
- Start from scatterplot
- Consider grid on scatterplot
- Define mutual information of empirical distribution on grid I_G
  - KL divergence of factored distribution from actual joint
  - Always ≥ 0
  - Information-theoretic measure of dependence; compression interpretation



From Figure 1 of Reshef et al. 2011

Technical definition, continued

- Now, fix grid size (x, y)
- Maximize I_G over grid layouts $\rightarrow I_G^*$
- Normalize to $M_{x,y} = \frac{I_G^*}{\log \min\{x,y\}}$
- Maximize again over (x, y) s.t. $x, y < B(n) \rightarrow M^*$
- M^* is MIC for pair of variables



From Figure 1 of Reshef et al. 2011

Computation, briefly

Hard to do this maximization

- Approximate search methods needed
- Dynamic-programming based solution
- Quite fast

Properties

MIC, as defined:

- Symmetric (from MI symmetry)
- $\rightarrow 0$ iff variables independent (with $B(n)$ conditions)
- $\rightarrow 1$ for functionally related variables
- Lower bound linked to R^2 for noisy functional relationships

Initial statistical reaction

That sounds great

Initial statistical reaction

That sounds great
But it can't be a panacea

Initial statistical reaction

That sounds great
But it can't be a panacea

- Must have lower power than, e.g., F-test for linear
- Nonfunctional $\rightarrow$ multimodal predictive distribution; harder than nonparametric regression
- Huge multiple comparisons problem

Initial statistical reaction

That sounds great
But it can't be a panacea

- Must have lower power than, e.g., F-test for linear
- Nonfunctional $\rightarrow$ multimodal predictive distribution; harder than nonparametric regression
- Huge multiple comparisons problem

And we have theorems

Outline

There's always a tuning parameter

Nonparametric techniques nearly always have smoothness parameters

- Kernel width, number of knots, penalty weight, etc.
- Require careful attention to ensure validity and efficiency

There's always a tuning parameter

Nonparametric techniques nearly always have smoothness parameters

- Kernel width, number of knots, penalty weight, etc.
- Require careful attention to ensure validity and efficiency

Here, it's grid size $B(n)$

- Large $B(n)$ $\rightarrow$ overfitting; find structure in everything
- Small $B(n)$ $\rightarrow$ oversmoothing; miss noisy/subtle structure

Pathological cases & overfitting

- Showed that $B(n) = \Omega(n^{1+\varepsilon})$, $\varepsilon > 0 \Rightarrow M^* \to 1$ almost surely
- So, $B(n)$ too large does overfit
- If $B(n) = O(n^{1-\varepsilon})$, $\varepsilon > 0$, MIC converges to correct value
- In particular, this implies MIC $\to 0$ for independent RVs

Choice of $B(n)$ — published method

Selected $B(n)$ via simulation in paper

- Showed $B(n) = n^{1-\varepsilon}$ had proper limits under independence
- Settled on $B(n) = n^{0.6}$
- Rationale not apparent; no power or predictive checks

What about the coefficient?

Usually need both rate and coefficient for smoothness parameters

- Standard to get both in nonparametric statistics
- Rates analytically, coefficient estimated/approximated
- Neither completely handled here
- Could compromise power

Outline

Simulations

Simon and Tibshirani addressed power concerns directly

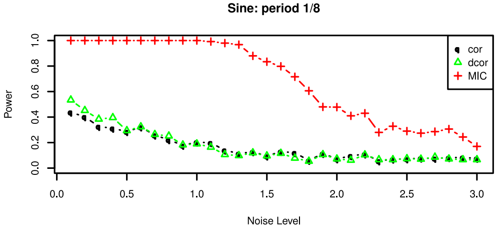
- Simulated from range of relationships with Gaussian noise
- Varied noise scale over factor of 3
- Evaluated frequentist power at FPR of 0.05
- Compared to Pearson and Brownian distance correlation

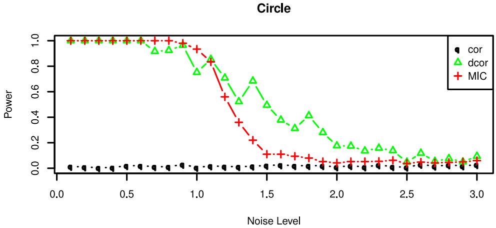
Brownian distance correlation

- Published by Székely and Rizzo in AoAS (2009)
- Uses distances between points and Brownian process approx
- Tuning parameter is power on distance
- Easy to compute (`energy` R package)

Power comparisons

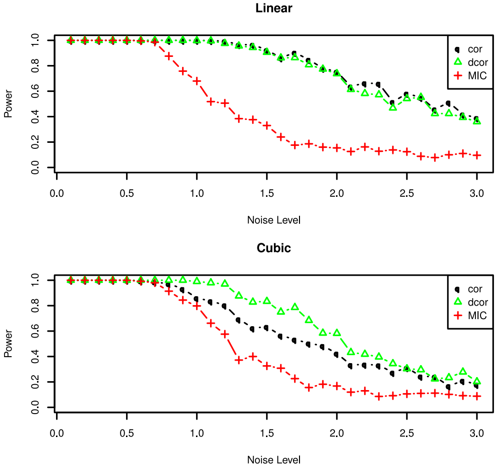
Alright for short-period sine wave and circular





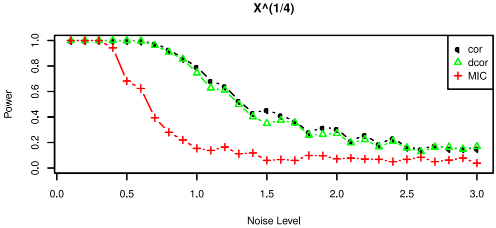
Power comparisons, continued

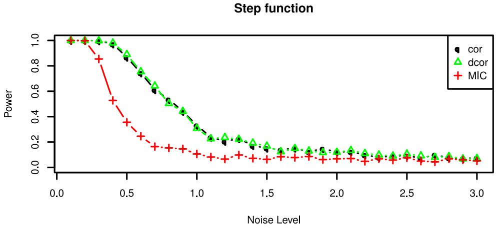
Underpowered for linear and cubic, as expected



Power comparisons, continued

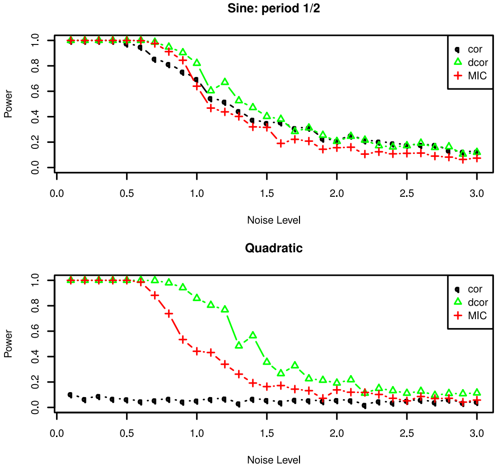
Surprisingly poor for $X^{1/4}$ and step functions





Power comparisons, continued

Alright, but not dominant, for long-period sine and quadratic



Discussion

As expected, there's no free lunch here

- Model-free method means less power for MIC
- Looking for extremely general forms of structure; inevitable tradeoffs
- Distance correlation is surprisingly good

Outline

Note

Concerns here are not particular to the Reshef et al. paper.

Note

Concerns here are not particular to the Reshef et al. paper. However, it does raise some interesting questions on this overall direction of research.

Pitfalls & potential of broader approach

Searching a vast amount of raw data for complex relationships can be problematic

- Often find mainly artifacts of the measurement process
- Conversely, using preprocessed data can show effects of processing rather than science
- Discovery is good goal, but is this too general?
  - Semi-supervised approaches
  - Hierarchical methods

Beyond bivariate

What types of complexity matter most?

- Increasing number of variables vs. increasing complexity
- Ideally both, but curse of dimensionality stings
- Often observe greater gains from covariates than complex low-dimensional structure
- Depends upon setting, of course

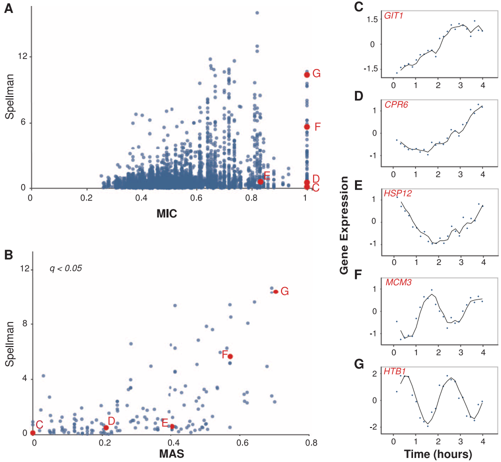
Independent detection vs. pooling information

Need to consider tradeoffs depending on richness of data per variable

- Little lost working independently with many data per variable
- With few observations per variable, pooling becomes more important
- Appears relevant even for some examples in paper (Spellman et al. data)

Example — Spellman data

Could benefit from hierarchical modeling



From Figure 5 of Reshef et al. 2011

Next steps with discovery-oriented analyses

Exploration and discovery, then ?

- After exploration phase, want stronger scientific results
- Predictive models, mechanistic hypotheses, etc.
- Dangers of inference with detected variables
- Distinction between EDA and data reduction
- Keeping sight of core modeling challenges

Location and publication

Where should statistics research appear?

- Nature/Science vs. statistics journals
- MIC & power law papers (Science)
- Contrast with FDR development (Jeff Leek's comments)