Discovering Potential Correlations via Hypercontractivity

Hyeji Kim*, Weihao Gao*, Sreeram Kannan+, Sewoong Oh*, Pramod Viswanath*

University of Illinois at Urbana-Champaign*, University of Washington+

INTRODUCTION

Discovering associations in large datasets

Example: Data for 300 indicators for 200 countries

Which pairs of indicators are associated?

~ 900,000 pairs of indicators!

Associations are used to make policy decisions

Important both in industry and scientific research

	Population	Energy Use	...	CO_2 Emissions
Afghanistan	26088	470	...	0.02
Albania	3172	761	...	0.98
...	...	...	...	...
Zambia	11696	620	...	0.21
Zimbabwe	13228	741	...	0.94

World Health Organization (WHO)

BACKGROUND

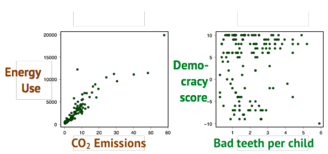
Correlation analysis to discover associations

Estimate correlation coefficients for all pairs of indicators

Pairs w. large corr coeff: candidates for strong association

dataset → Correlation Estimator →

X	Y	Cor(X,Y)
Democracy score	Bad teeth per child	0.02
HIV	Deaths from HIV	0.96
CO2 Emission	Energy Use	0.8
...	...	...
Child per woman	Fertility rate	0.999

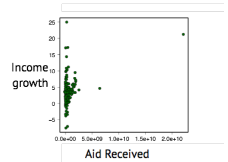


Examples of strong (left) and week (right) associations from WHO dataset

PROBLEM STATEMENT

Motivation:

Existing correlation estimators discover average correlations but fail to discover *potential* correlations



Discovering potential correlations can *affect policy decisions* and *lead to scientific findings*

Goal: discover potential correlations

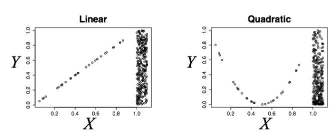
METHOD

1. Propose 7 axioms for a measure of potential correlation

$0 \le \rho(X,Y) \le 1$

$\rho(X,Y) = 0$ iff X and Y are independent

$\boxed{\rho(X,Y) = 1 \text{ if } Y = f(X) \text{ for } (X,Y) \in \mathcal{X}_r \times \mathcal{Y} \text{ for some } \mathcal{X}_r \subseteq \mathcal{X}}$



2. Show *hypercontractivity coefficient* satisfies all axioms

$$s(X;Y) \equiv \sup_{U-X-Y} \frac{I(U;Y)}{I(U;X)}$$

3. Propose a novel estimator

- based on an alternative definition

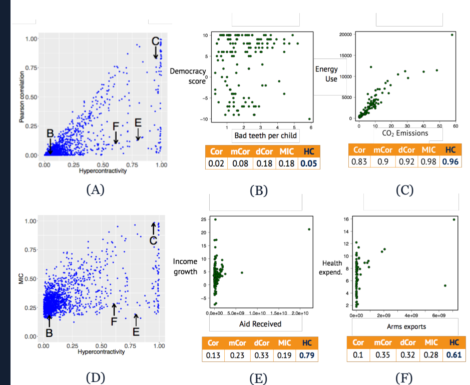
$$s(X;Y) = \sup_{r(x)\neq p(x)} \frac{D_{\mathrm{KL}}(r(y)\|p(y))}{D_{\mathrm{KL}}(r(x)\|p(x))}$$

$$\text{where } r(y) = \sum_x r(x)p(y|x)$$

- via joint optimization and estimation

EXPERIMENTS

1. WHO dataset



(A): Scatter plot of Pearson correlation vs. HC

(D): Scatter plot of Maximal Info. Coefficient vs. HC

(B): All correlations are small

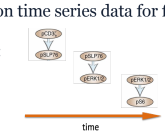
(C): All correlations are large

(E) and (F): Only HC discovers potential correlations

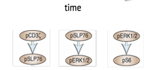
2. Genetic Pathway Recovery

Gene expression time series data for four genes

Biological fact:



If we only know



can we recover the sequential order of influence?

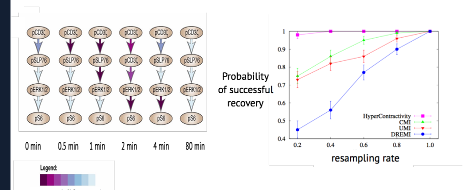


Figure:Krishnaswamy et al. 2014. Science

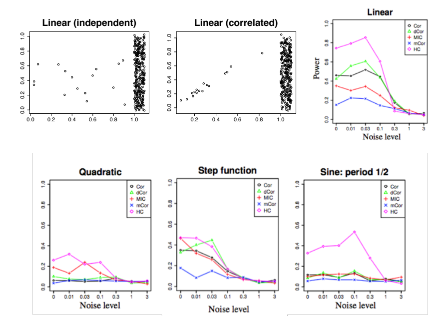
Hypercontractivity: robust measure strength of influence

EXPERIMENTS

3. Power test

Binary hypothesis testing of potential correlations

Power: true positive rate for a fixed false positive rate



HC is more powerful than others in hypothesis testing of canonical examples of potential correlations

CONCLUSION

1. We postulate a set of natural axioms that we expect a measure of potential correlation to satisfy
2. We show that *rate* of information bottleneck, i.e., the *hypercontractivity* coefficient (HC), satisfies all the proposed axioms
3. We provide a novel estimator for HC
4. Experimental results: WHO datasets, genetic pathway recovery, power tests

ACKNOWLEDGEMENTS

This work was partially supported by NSF grants CNS-1527754, CNS-1718270, CCF-1553452, CCF-1617745, CCF-1651236, CCF-1705007, and GOOGLE Faculty Research Award.

ILLINOIS

UNIVERSITY of WASHINGTON