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

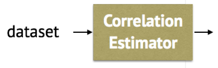
World Health Organization (WHO)

BACKGROUND

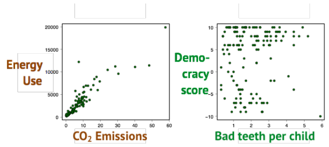
Correlation analysis to discover associations

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

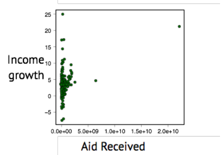


Examples of strong (left) and week (right) associations from WHO dataset

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

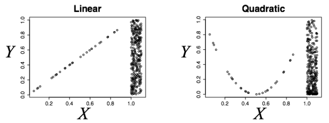
METHOD

1. Propose 7 axioms for a measure of potential correlation

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

$\rho(X,Y) = 0$ 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

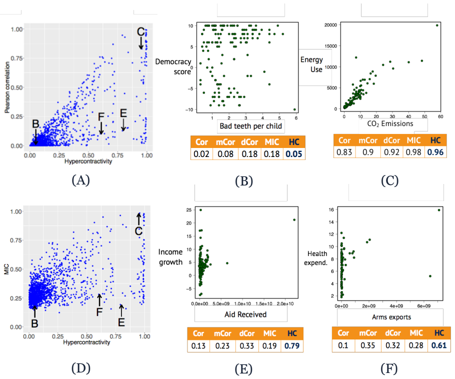
- based on an alternative definition

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

- via joint optimization and estimation

EXPERIMENTS

1. WHO dataset



(B)

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	0.05

(C)

Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	0.96

(E)

Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	0.79

(F)

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	0.61

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

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

(B): All correlations are small

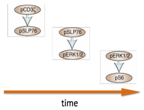
(C): All correlations are large

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

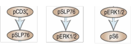
2. Genetic Pathway Recovery

Gene expression time series data for four genes

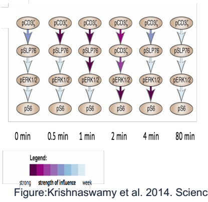
Biological fact:



If we only know



can we recover the sequential order of influence?



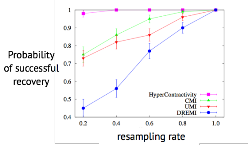


Figure:Krishnaswamy et al. 2014. Science

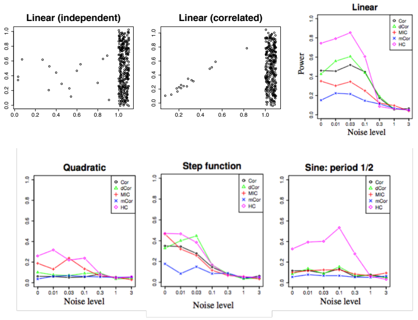
Hypercontractivity: robust measure strength of influence

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

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!

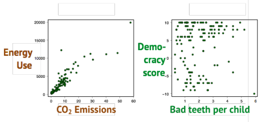
? ?

	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)

Correlation analysis to discover associations

Estimate correlation coefficients for all pairs of indicators



X	Y	Cor(X,Y)
Democracy score	Bad teeth per child	0.2
HIV	Deaths from HIV	0.96
CO2 Emission	Energy Use	0.8
...	...	...
Child per woman	Fertility rate	0.999

Rank pairs according to correlation coefficients

Candidates for strong associations

Rank	X	Y	Cor(X,Y)
1	Child per woman	Fertility rate	0.999
2	HIV	Deaths from HIV	0.96
3	CO2 Emission	Energy Use	0.8
...	...	...	...
300* 300	Democracy score	Bad teeth per child	0.2

Different correlation estimators discover diff. associations

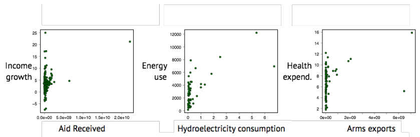
Pearson Correlation Estimator E[XY] discovers linear associations

Maximal Correlation Estimator $\max_{f,g} E[f(X)g(Y)]$ discovers functional associations

PROBLEM STATEMENT

Motivation:

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



Examples of potential correlations from WHO dataset

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

Goal:

Discover potential correlations

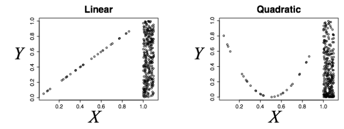
Our Approach:

1. Postulate 7 axioms for a measure of potential correlation, including

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

$\rho(X,Y) = 0$ 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 equivalent definition

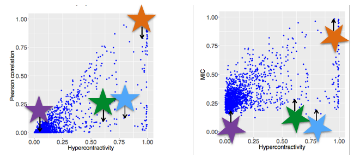
$$s(X;Y) = \sup_{r(x)\neq p(x)} \frac{D_{\mathrm{KL}}(r(y)\|p(y))}{D_{\mathrm{KL}}(r(x)\|p(x))}$$
$$\text{where} \quad r(y) = \sum_x r(x)p(y|x)$$

- joint optimization and estimation

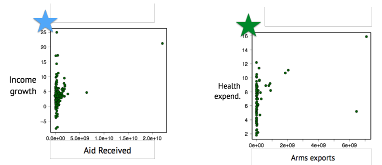
EXPERIMENTS

1. WHO dataset

Scatter plots of Pearson correlation vs. HC (left) and MIC vs. HC (right)



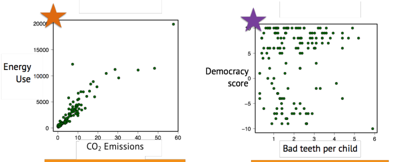
Only hypercontractivity discovers potential correlations



Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	**0.79**

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	**0.61**

Hypercontracitivy & others discover average correlations



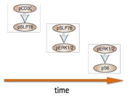
Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	**0.96**

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	**0.05**

2. Genetic Pathway Recovery

Gene expression time series data for four genes

Biological fact:



If we only know

pCD3ζ
pSLP76
pERK1/2
pS6

can we recover the sequential order of influence?

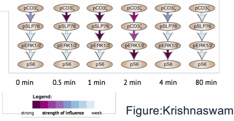
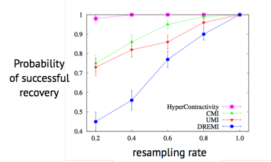


Figure:Krishnaswamy et al. 2014. Science

EXPERIMENTS

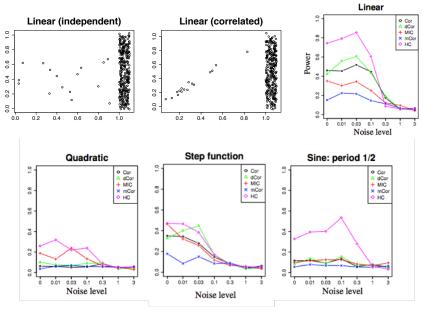
Hypercontractivity: robust measure strength of influence



3. Power test

Binary hypothesis testing of potential correlation

Power: true positive rate for a fixed false positive rate



HC is more powerful than others for 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

ILLINOIS

UNIVERSITY *of* WASHINGTON

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

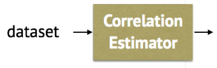
World Health Organization (WHO)

BACKGROUND

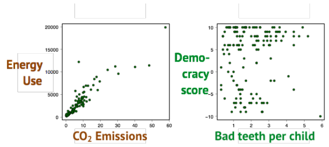
Correlation analysis to discover associations

Estimate correlation coefficients for all pairs of indicators

Pairs w. large corr coeff: candidates for strong association



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



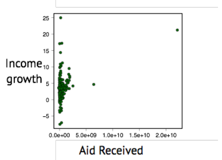
Examples of strong (left) and week (right) associations from WHO dataset

PROBLEM STATEMENT

Motivation:

All correlation estimators discover average correlations

Fail to discover *potential* correlation



Discovering potential correlations can

- affect policy decisions
- lead to scientific findings

Goal:

Discover potential correlations

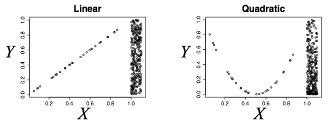
Our approach:

1. Propose 7 axioms for a measure of potential correlation

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

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

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

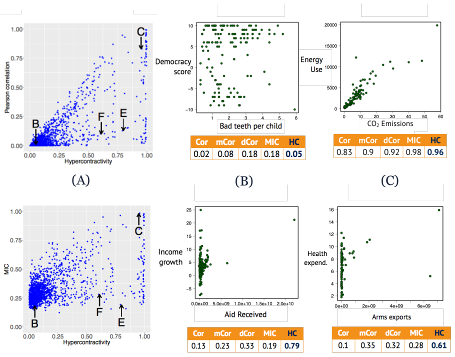
$$s(X;Y) = \sup_{r(x) \neq p(x)} \frac{D_{\mathrm{KL}}(r(y) \| p(y))}{D_{\mathrm{KL}}(r(x) \| p(x))}$$

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

- via joint optimization and estimation

EXPERIMENTS

1. WHO dataset



(B)

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	**0.05**

(C)

Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	**0.96**

(E)

Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	**0.79**

(F)

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	**0.61**

(A) (B) (C) (D) (E) (F)

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

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

(B): All correlations are small

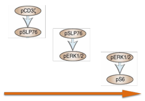
(C): All correlations are large

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

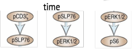
2. Genetic Pathway Recovery

Gene expression time series data for four genes

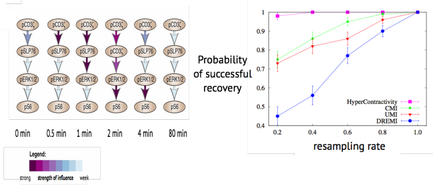
Biological fact:



If we only know



can we recover the sequential order of influence?



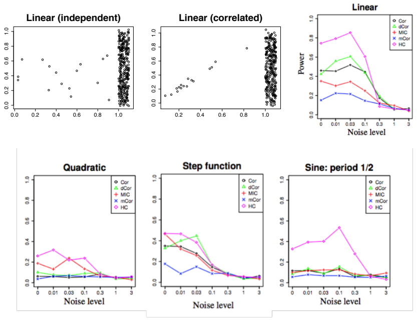
Hypercontractivity: robust measure strength of influence

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

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?

Associations are used to make policy decisions

Important both in industry and scientific research

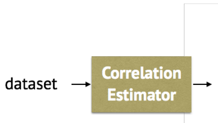
	Population	Energy Use	...	CO_2 Emissions
Afghanistan	26088	470	...	0.02
Albania	3172	761	...	0.98
...	...	...	...	...
Zambia	11696	620	...	0.21
Zimbabwe	13228	741	...	0.94

World Health Organization (WHO)

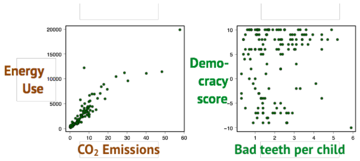
CORRELATION ANALYSIS

Estimate correlation coefficients for all pairs of indicators

Correlation coefficients: measure strength of association



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



Different correlation estimators discover diff. associations

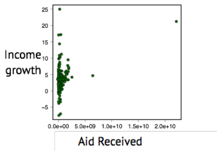
Example: Pearson correlations – linear associations

Maximal correlations – Functional associations

POTENTIAL CORRELATION

Goal

Discover potential correlation from large datasets



Problem

All correlation estimators discover average correlations

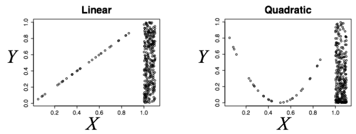
Fail to discover *potential* correlation

Our Approach

Provide a measure of potential correlation and estimator

(1) Propose axioms for a measure of potential correlation

$$0 \le \rho(X,Y) \le 1$$
$$\rho(X,Y) = 0 \text{ iff } X \text{ and } Y \text{ are independent}$$
$$\rho(X,Y) = 1 \text{ if } Y = f(X) \text{ for } (X,Y) \in \mathcal{X}_r \times \mathcal{Y} \text{ for some } \mathcal{X}_r \subseteq \mathcal{X}$$



(2) Hypercontractivity coefficient satisfies all axioms

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

(3) Estimator for Hypercontractivity coefficient (HC)

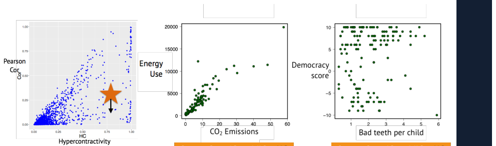
$$s(X;Y) = \sup_{r(x)\neq p(x)} \frac{D_{\mathrm{KL}}(r(y)\|p(y))}{D_{\mathrm{KL}}(r(x)\|p(x))}$$

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

- Joint optimization and estimation

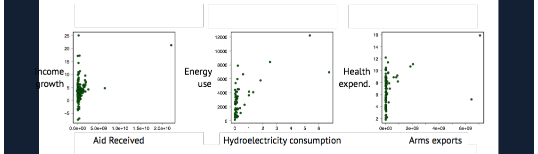
EXPERIMENTS

(1) WHO dataset



Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	**0.96**

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	**0.05**



Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	**0.79**

Cor	mCor	dCor	MIC	HC
0.42	0.55	0.61	0.29	**0.91**

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	**0.61**

(A) and (D): Scatter plot of correlation measures

(B): All correlations are small

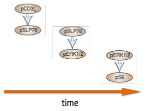
(C): All correlations are large

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

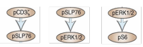
(2) Genetic Pathway Recovery

Gene expression time series data for four genes

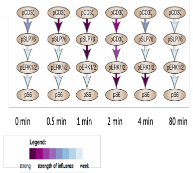
Biological fact:

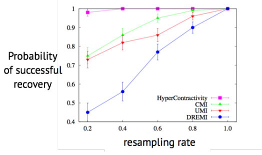


If we only know



can we recover the sequential order of influence?



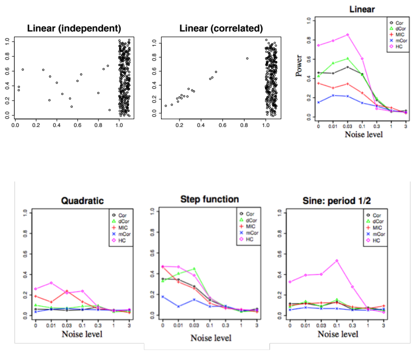


Hypercontractivity: robust measure strength of influence

EXPERIMENTS

(3) Power test

Binary hypothesis testing of potential correlation



CONCLUSION

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

   WHO datasets, genetic pathway recovery, power tests

ACKNOWLEDGEMENTS

Check to make sure you’ve acknowledged partner and funding agencies, either with text or with their logos.

ILLINOIS

UNIVERSITY of WASHINGTON

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	...	CO₂ Emissions
Afghanistan	26088	470	...	0.02
Albania	3172	761	...	0.98
...	...	...	...	...
Zambia	11696	620	...	0.21
Zimbabwe	13228	741	...	0.94

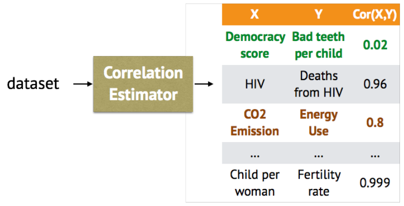
World Health Organization (WHO)

BACKGROUND

Correlation analysis to discover associations

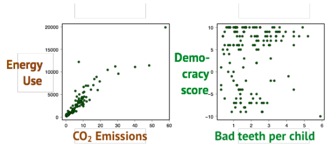
Estimate correlation coefficients for all pairs of indicators

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



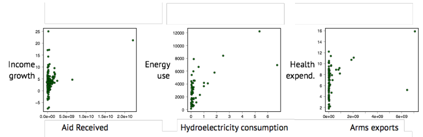
Examples of strong (left) and week (right) associations from WHO dataset

PROBLEM STATEMENT

Motivation

Exist. correlation estimators discover average correlation

Fail to discover *potential* correlation



Examples of potential correlations from WHO dataset

Discovering potential correlations between aid received vs. income growth: affect policy decisions

Goal: discover potential correlation

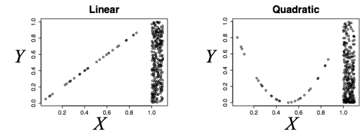
METHOD

1. Propose axioms for a measure of potential correlation

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

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

$\rho(X,Y) = 1$ if $Y = f(X)$ for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$ 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 for hypercontractivity coeff.

- based on equivalent definition of s(X;Y):

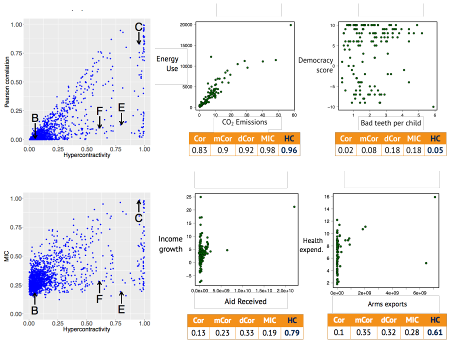
$$s(X;Y) = \sup_{r(x)\neq p(x)} \frac{D_{\mathrm{KL}}(r(y)\|p(y))}{D_{\mathrm{KL}}(r(x)\|p(x))}$$

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

- joint optimization and estimation

EXPERIMENTS

(1) WHO dataset



Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	0.96

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	0.05

Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	0.79

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	0.61

(A) and (D): Scatter plot of correlation measures

(B): All correlations are small

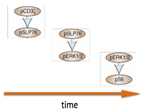
(C): All correlations are large

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

(2) Genetic Pathway Recovery

Gene expression time series data for four genes

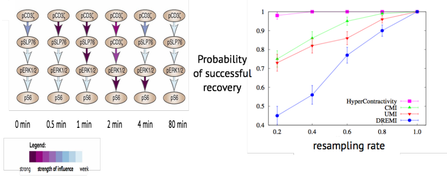
Biological fact:



If we only know

pCD3ζ
pSLP76
pSLP76
pERK1/2
pERK1/2
pS6

can we recover the sequential order of influence?



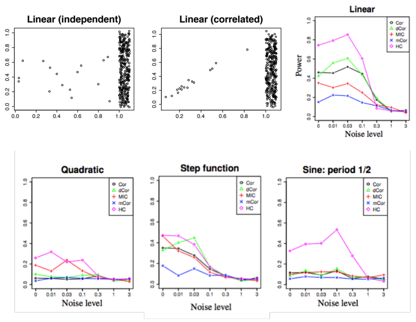
Hypercontractivity: robust measure strength of influence

EXPERIMENTS

(3) Power test

Binary hypothesis testing of potential correlation

Power: true positive rate for a fixed false positive rate



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

Check to make sure you' ve acknowledged partner and funding agencies, either with text or with their logos.

ILLINOIS

UNIVERSITY of WASHINGTON

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

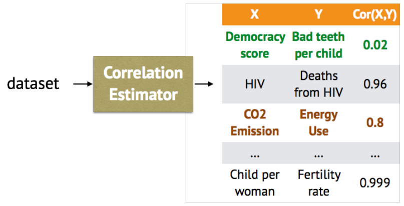
World Health Organization (WHO)

BACKGROUND

Correlation analysis to discover associations

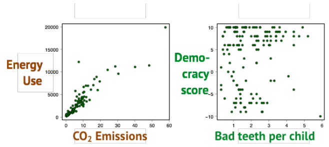
Estimate correlation coefficients for all pairs of indicators

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



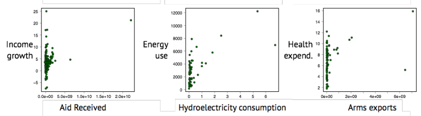
Examples of strong (left) and week (right) associations from WHO dataset

PROBLEM STATEMENT

Motivation

Exist. correlation estimators discover average correlation

Fail to discover *potential* correlation



Examples of potential correlations from WHO dataset

Discovering potential correlations between aid received vs. income growth: affect policy decisions

Goal: discover potential correlation

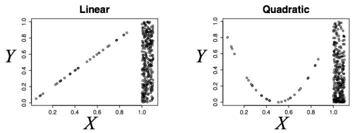
METHOD

1. Propose axioms for a measure of potential correlation

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

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

$\rho(X,Y) = 1$ if $Y = f(X)$ for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$ 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 for hypercontractivity coeff.

- based on equivalent definition of s(X;Y):

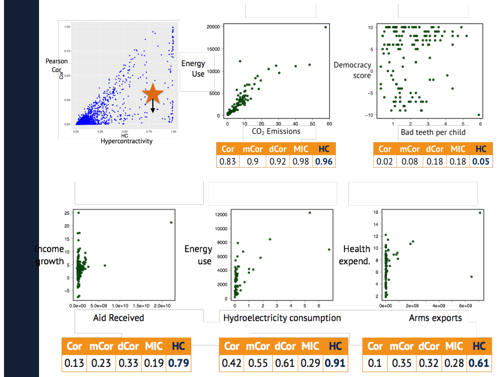
$$s(X;Y) = \sup_{r(x)\neq p(x)} \frac{D_{\mathrm{KL}}(r(y)\|p(y))}{D_{\mathrm{KL}}(r(x)\|p(x))}$$

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

- joint optimization and estimation

EXPERIMENTS

(1) WHO dataset



Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	**0.96**

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	**0.05**

Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	**0.79**

Cor	mCor	dCor	MIC	HC
0.42	0.55	0.61	0.29	**0.91**

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	**0.61**

(A) and (D): Scatter plot of correlation measures

(B): All correlations are small

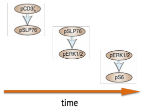
(C): All correlations are large

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

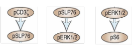
(2) Genetic Pathway Recovery

Gene expression time series data for four genes

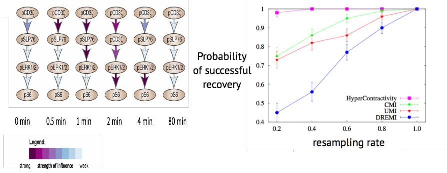
Biological fact:



If we only know



can we recover the sequential order of influence?



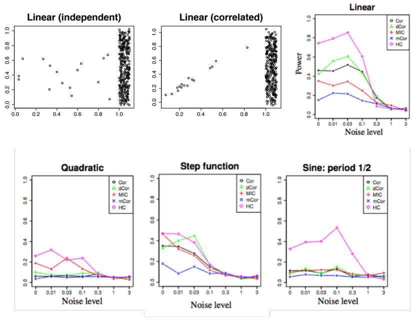
Hypercontractivity: robust measure strength of influence

EXPERIMENTS

(3) Power test

Binary hypothesis testing of potential correlation

Power: true positive rate for a fixed false positive rate



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

Check to make sure you' ve acknowledged partner and funding agencies, either with text or with their logos.

ILLINOIS

UNIVERSITY of WASHINGTON

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)

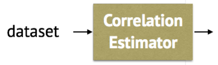
CORRELATION ANALYSIS

Correlation analysis to discover associations

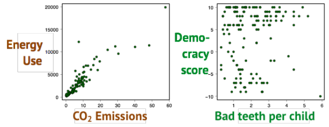
Correlation coefficient: a measure to quantify association

Estimate correlation coefficients for all pairs of indicators

Pairs w. large corr coeff: candidates for strong association



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

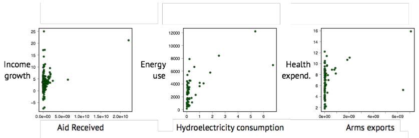


Examples of strong (left) and week (right) associations from WHO dataset

POTENTIAL CORRELATION

Goal

Discover *potential correlation* from large datasets



Examples of potential correlations from WHO dataset

Problem

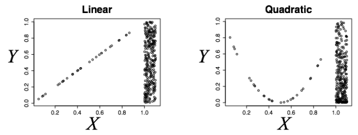
Exist. correlation estimators discover average correlation

Fail to discover *potential* correlation

Our Approach

1. Propose axioms for a measure of potential correlation

$0 \leq \rho(X,Y) \leq 1$
$\rho(X,Y) = 0$ iff X and Y are independent
$\rho(X,Y) = 1$ if $Y = f(X)$ for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$ 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 for HC

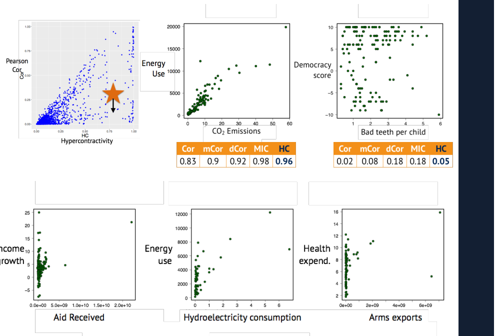
- based on equivalent definition of s(X;Y):

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

- joint optimization and estimation

EXPERIMENTS

(1) WHO dataset



Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	0.96

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	0.05

Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	0.79

Cor	mCor	dCor	MIC	HC
0.42	0.55	0.61	0.29	0.91

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	0.61

(A) and (D): Scatter plot of correlation measures

(B): All correlations are small

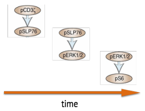
(C): All correlations are large

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

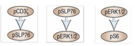
(2) Genetic Pathway Recovery

Gene expression time series data for four genes

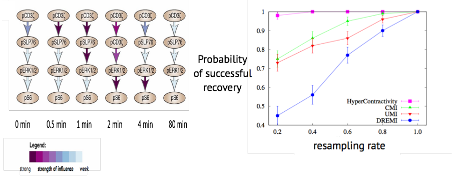
Biological fact:



If we only know



can we recover the sequential order of influence?



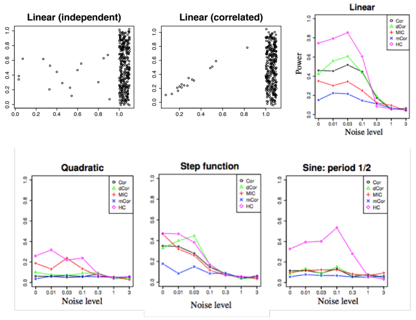
Hypercontractivity: robust measure strength of influence

EXPERIMENTS

(3) Power test

Binary hypothesis testing of potential correlation

Power: true positive rate for a fixed false positive rate



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

Check to make sure you' ve acknowledged partner and funding agencies, either with text or with their logos.

ILLINOIS

UNIVERSITY of WASHINGTON

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	...	CO₂ Emissions
Afghanistan	26088	470	...	0.02
Albania	3172	761	...	0.98
...	...	...	...	...
Zambia	11696	620	...	0.21
Zimbabwe	13228	741	...	0.94

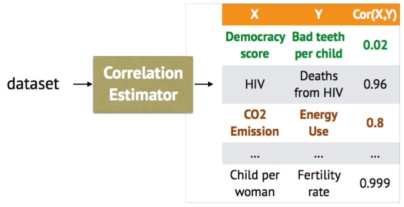
World Health Organization (WHO)

Correlation analysis to discover associations

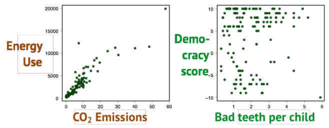
Correlation coefficient: a measure to quantify association

Estimate correlation coefficients for all pairs of indicators

Pairs w. large corr coeff: candidates for strong association



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

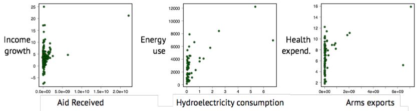


Examples of strong (left) and week (right) associations from WHO dataset

POTENTIAL CORRELATION

Goal

Discover *potential correlation* from large datasets



Examples of potential correlations from WHO dataset

Problem

Exist. correlation estimators discover average correlation

Fail to discover *potential* correlation

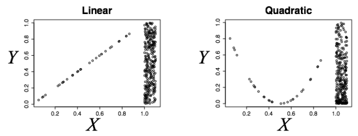
Our Approach

1. Propose axioms for a measure of potential correlation

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

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

$\rho(X,Y) = 1$ if $Y = f(X)$ for $(X,Y) \in \mathcal{X}_r \times \mathcal{Y}$ 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 for HC

- based on equivalent definition of s(X;Y):

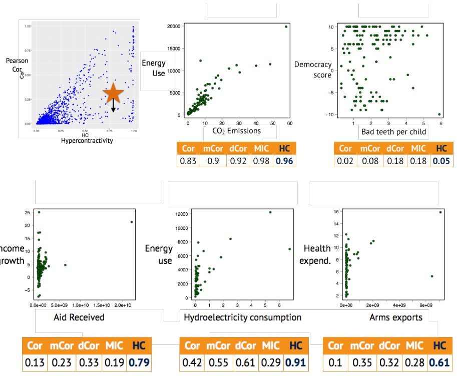
$$s(X;Y) = \sup_{r(x)\neq p(x)} \frac{D_{\mathrm{KL}}(r(y)\|p(y))}{D_{\mathrm{KL}}(r(x)\|p(x))}$$

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

- joint optimization and estimation

EXPERIMENTS

(1) WHO dataset



Cor	mCor	dCor	MIC	HC
0.83	0.9	0.92	0.98	**0.96**

Cor	mCor	dCor	MIC	HC
0.02	0.08	0.18	0.18	**0.05**

Cor	mCor	dCor	MIC	HC
0.13	0.23	0.33	0.19	**0.79**

Cor	mCor	dCor	MIC	HC
0.42	0.55	0.61	0.29	**0.91**

Cor	mCor	dCor	MIC	HC
0.1	0.35	0.32	0.28	**0.61**

(A) and (D): Scatter plot of correlation measures

(B): All correlations are small

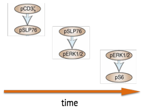
(C): All correlations are large

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

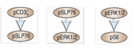
(2) Genetic Pathway Recovery

Gene expression time series data for four genes

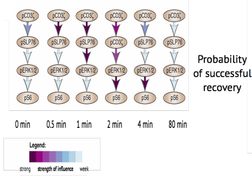
Biological fact:

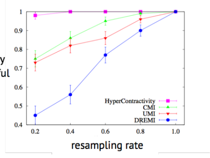


If we only know



can we recover the sequential order of influence?





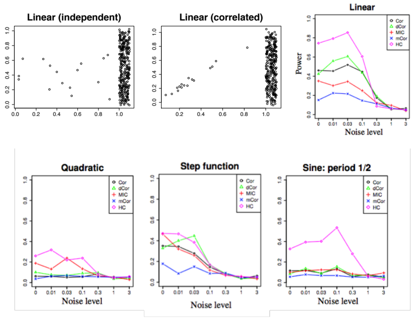
Hypercontractivity: robust measure strength of influence

EXPERIMENTS

(3) Power test

Binary hypothesis testing of potential correlation

Power: true positive rate for a fixed false positive rate



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

Check to make sure you' ve acknowledged partner and funding agencies, either with text or with their logos.

ILLINOIS

W UNIVERSITY *of* WASHINGTON