Generalized R^2 Measures for a Mixture of Bivariate Linear Dependences

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

Joint work with Drs. Xin Tong (USC) and Peter J. Bickel (UC Berkeley)

Motivation: Maximal Information Coefficient

Science Home News Journals Topics Careers

SHARE

RESEARCH ARTICLE

Detecting Novel Associations in Large Data Sets

David N. Reshef[1,2,3,*,†], Yakir A. Reshef[2,4,*,†], Hilary K. Finucane[5], Sharon R. Grossman[2,6], Gilean McVean[3,7], Peter J. Turnb...

+ See all authors and affiliations

Science 16 Dec 2011:
Vol. 334, Issue 6062, pp. 1518-1524
DOI: 10.1126/science.1205438

0

SHARE

PERSPECTIVE | MATHEMATICS

A Correlation for the 21st Century

Terry Speed

+ See all authors and affiliations

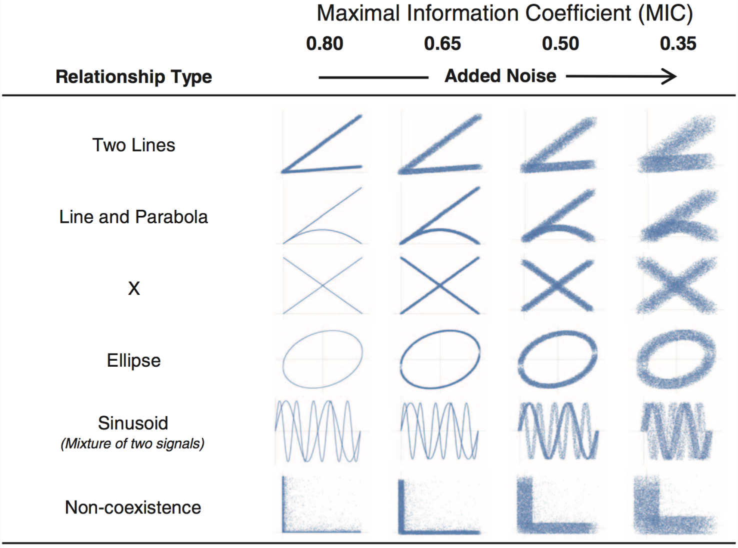
Science 16 Dec 2011:
Vol. 334, Issue 6062, pp. 1502-1503
DOI: 10.1126/science.1215894

Motivation: Maximal Information Coefficient

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

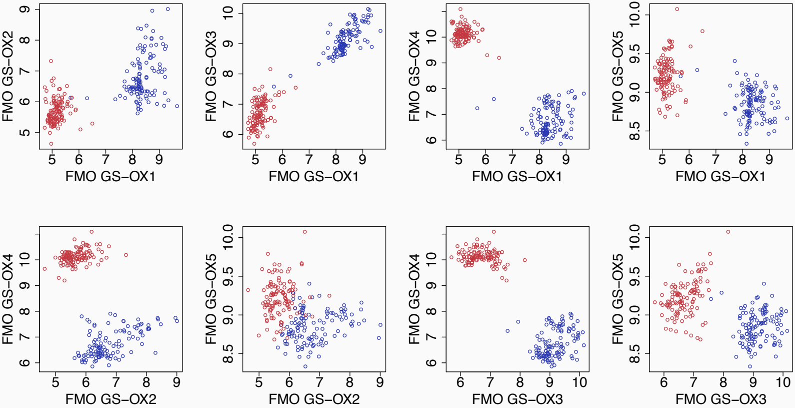
These maximal correlation values < 1 were due to lack of convergence

Motivation: Maximal Information Coefficient



Are these “non-functional” patterns important?

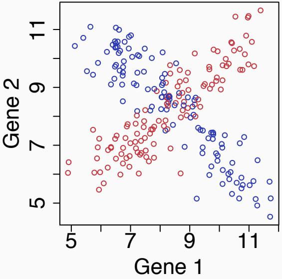
Motivation: Gene Expression Analysis



Five functionally related genes in *A. thaliana* (Kim et al., 2012)

Red: root tissues; Blue: shoot tissues

Motivation: Simpson's Paradox

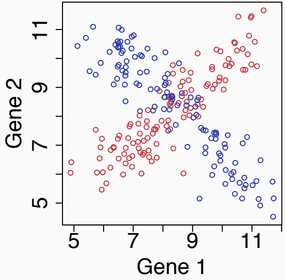


Pearson cor (red) ≈ 0.8

Pearson cor (blue) ≈ -0.8

Pearson cor (all) $\approx$ 0

Motivation: Simpson's Paradox



Pearson cor (red) ≈ 0.8

Pearson cor (blue) ≈ -0.8

Pearson cor (all) ≈ 0

Science Home News Journals Topics Careers

SHARE

ARTICLES

Sex Bias in Graduate Admissions: Data from Berkeley

P. J. Bickel[1], E. A. Hammel[1], J. W. O'Connell[1]

+ See all authors and affiliations

Science 07 Feb 1975:
Vol. 187, Issue 4175, pp. 398-404
DOI: 10.1126/science.187.4175.398

0

Review: Scalar-valued Association Measures

Measure: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Relationship Type		Measure
Functional (1-to-1)	Linear	Pearson correlation
	Monotone	Spearman's rank correlation Kendall's τ
	General	maximal correlation (Rényi, 1959) correlation curves (Bjerve and Doksum, 1993) principal curves (Delicado and Smrekar, 2009) generalized measures of correlation (Zheng et al., 2012) count statistics (Wang et al., 2014) G^2 statistic (Wang et al., 2017)
Dependent		Hoeffding's D mutual information HSIC (Gretton et al., 2005) distance correlation (Székely et al., 2007) maximal information coefficient (Reshef et al., 2011) HHG association test statistic (Heller et al., 2012)

Review: Scalar-valued Association Measures

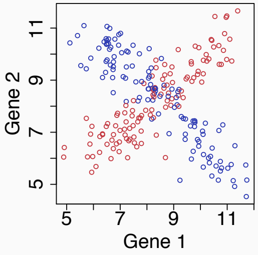
Measure: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Measures for Relationship Type		Interpretability	Flexibility
Functional (1-to-1)	Linear	best	worst
	Monotone	↓	↑
	General		
Dependent		worst	best

Review: Scalar-valued Association Measures

Measure: $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$

Measures for Relationship Type		Interpretability	Flexibility
Functional (1-to-1)	Linear	best ↓	worst ↑
	Monotone		
	General		
Dependent		worst	best



Mixture of linear dependences

- Widespread
- Easy to interpret
- Calling for a new powerful measure

Review: Mixture of Linear Dependences

Over 40 years

- Statistics
- Economics
- Social sciences
- Machine learning

Model parameter estimation & inference:

- (Quandt and Ramsey, 1978; De Veaux, 1989)
- (Jacobs et al., 1991; Jones and McLachlan, 1992)
- (Wedel and DeSarbo, 1994; Turner, 2000)
- (Hawkins et al., 2001; Hurn et al., 2003)
- (Leisch, 2008; Benaglia et al., 2009)
- (Scharl et al., 2009)

Algorithm:

- (Murtaph and Raftery, 1984)

Review: Mixture of Linear Dependences

Over 40 years

- Statistics
- Economics
- Social sciences
- Machine learning

Model parameter estimation & inference:

- (Quandt and Ramsey, 1978; De Veaux, 1989)
- (Jacobs et al., 1991; Jones and McLachlan, 1992)
- (Wedel and DeSarbo, 1994; Turner, 2000)
- (Hawkins et al., 2001; Hurn et al., 2003)
- (Leisch, 2008; Benaglia et al., 2009)
- (Scharl et al., 2009)

Algorithm:

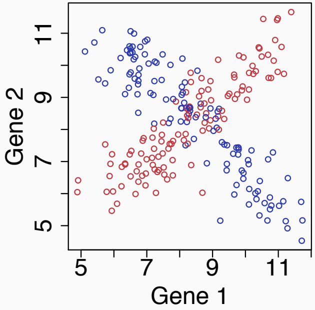
- (Murtaph and Raftery, 1984)

Association measure: question of interest

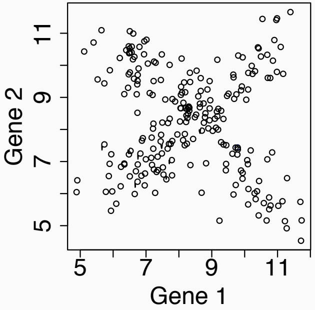
Formulation: Supervised and Unsupervised Scenarios

- $X, Y \in \mathbb{R}$ — random variables whose relationship is of interest
  - observed
- $Z \in \{1, \ldots, K\}$ — indicator of linear relationship
  - observed (**supervised scenario**)
  - hidden (**unsupervised scenario**)
- When $K = 1$, only the supervised scenario exists

Supervised



Unsupervised



Supervised Population Generalized R^2: $\rho^2_{\mathcal{G}(\mathcal{S})}$

Given the joint distribution of (X, Y, Z), denote

$$p_{k(\mathcal{S})} := \mathbb{P}(Z = k)\,,\ k = 1, \ldots, K\,,\ \text{with} \sum_{k=1}^{K} p_{k(\mathcal{S})} = 1\,.$$

and

$$\rho_{k(\mathcal{S})} := \frac{\mathrm{cov}(X, Y | Z = k)}{\sqrt{\mathrm{var}(X|Z = k)}\sqrt{\mathrm{var}(Y|Z = k)}}$$

as the population Pearson correlation of $(X, Y)|Z = k$.

Definition: $\rho^2_{\mathcal{G}(\mathcal{S})}$

The supervised population generalized R^2 is defined as

$$\rho^2_{\mathcal{G}(\mathcal{S})} := \mathbb{E}_Z\left[\rho^2_{Z(\mathcal{S})}\right] = \mathbb{E}_Z\left[\frac{\mathrm{cov}^2(X, Y|Z)}{\mathrm{var}(X|Z)\mathrm{var}(Y|Z)}\right] = \sum_{k=1}^{K} p_{k(\mathcal{S})} \cdot \rho^2_{k(\mathcal{S})}$$

K-line Interpretation the Supervised Scenario

- Denote by $\beta = (a, b, c)^\mathsf{T}$ a **line**

$$\left\{(x, y)^\mathsf{T} : ax + by + c = 0, \text{ where } a, b, c \in \mathbb{R} \text{ with } a \neq 0 \text{ or } b \neq 0\right\} \subset \mathbb{R}^2$$

K-line Interpretation the Supervised Scenario

- Denote by $\beta = (a, b, c)^\mathsf{T}$ a **line**

  $$\left\{(x, y)^\mathsf{T} : ax + by + c = 0,\ \text{where } a, b, c \in \mathbb{R} \text{ with } a \neq 0 \text{ or } b \neq 0\right\} \subset \mathbb{R}^2$$

- **Perpendicular distance** between $(x, y)^\mathsf{T}$ and β is $d_\perp : \mathbb{R}^2 \times \mathbb{R}^3 \mapsto \mathbb{R}$:

  $$d_\perp\left((x, y)^\mathsf{T}, \beta\right) = \frac{|ax + by + c|}{\sqrt{a^2 + b^2}}$$

  Symmetric between x and y

K-line Interpretation the Supervised Scenario

- Denote by $\boldsymbol{\beta} = (a, b, c)^\mathsf{T}$ a **line**

$$\left\{(x, y)^\mathsf{T} : ax + by + c = 0, \text{ where } a, b, c \in \mathbb{R} \text{ with } a \neq 0 \text{ or } b \neq 0\right\} \subset \mathbb{R}^2$$

- **Perpendicular distance** between $(x, y)^\mathsf{T}$ and $\boldsymbol{\beta}$ is $d_\perp : \mathbb{R}^2 \times \mathbb{R}^3 \mapsto \mathbb{R}$:

$$d_\perp\left((x, y)^\mathsf{T}, \boldsymbol{\beta}\right) = \frac{|ax + by + c|}{\sqrt{a^2 + b^2}}$$

Symmetric between x and y

Definition: Supervised Population k-th Line Center

$$\boldsymbol{\beta}_{k(\mathcal{S})} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[d_\perp^2\left((X, Y)^\mathsf{T}, \boldsymbol{\beta}\right) \middle| Z = k\right]$$

K-line Interpretation the Supervised Scenario

Definition: Supervised Population k-th Line Center

$$\boldsymbol{\beta}_{k(\mathcal{S})} = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[d_{\perp}^2 \left((X, Y)^{\mathsf{T}}, \boldsymbol{\beta} \right) \middle| Z = k \right]$$

corresponds to the first principal component of

$$\boldsymbol{\Sigma}_{k(\mathcal{S})} := \begin{bmatrix} \mathrm{var}(X|Z=k) & \mathrm{cov}(X, Y|Z=k) \\ \mathrm{cov}(X, Y|Z=k) & \mathrm{var}(Y|Z=k) \end{bmatrix}$$

(Jolliffe, 2011)

$B_{K(\mathcal{S})} = \{\boldsymbol{\beta}_{1(\mathcal{S})}, \ldots, \boldsymbol{\beta}_{K(\mathcal{S})}\}$: supervised population line centers

Supervised Sample Generalized R^2: $R^2_{\mathcal{G}(\mathcal{S})}$

Consider a sample $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$

Definition: $R^2_{\mathcal{G}(\mathcal{S})}$

The supervised sample generalized R^2 is defined as

$$R^2_{\mathcal{G}(\mathcal{S})} := \sum_{k=1}^{K} \widehat{p}_{k(\mathcal{S})} \cdot \widehat{\rho}^2_{k(\mathcal{S})}$$

where

$$\widehat{p}_{k(\mathcal{S})} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Z_i = k)$$

$$\widehat{\rho}^2_{k(\mathcal{S})} := \frac{\left[\sum_{i=1}^{n} (X_i - \bar{X}_{k(\mathcal{S})})(Y_i - \bar{Y}_{k(\mathcal{S})}) \mathbb{I}(Z_i = k)\right]^2}{\left[\sum_{i=1}^{n} (X_i - \bar{X}_{k(\mathcal{S})})^2 \mathbb{I}(Z_i = k)\right] \left[\sum_{i=1}^{n} (Y_i - \bar{Y}_{k(\mathcal{S})})^2 \mathbb{I}(Z_i = k)\right]}$$

with

- $\bar{X}_{k(\mathcal{S})} = \frac{1}{n_{k(\mathcal{S})}} \sum_{i=1}^{n} X_i \mathbb{I}(Z_i = k)$; $\bar{Y}_{k(\mathcal{S})} = \frac{1}{n_{k(\mathcal{S})}} \sum_{i=1}^{n} Y_i \mathbb{I}(Z_i = k)$
- $n_{k(\mathcal{S})} = \sum_{i=1}^{n} \mathbb{I}(Z_i = k)$

Unsupervised Population Line Centers

Given the joint distribution of (X, Y)

Definition: $B_{K(\mathcal{U})}$

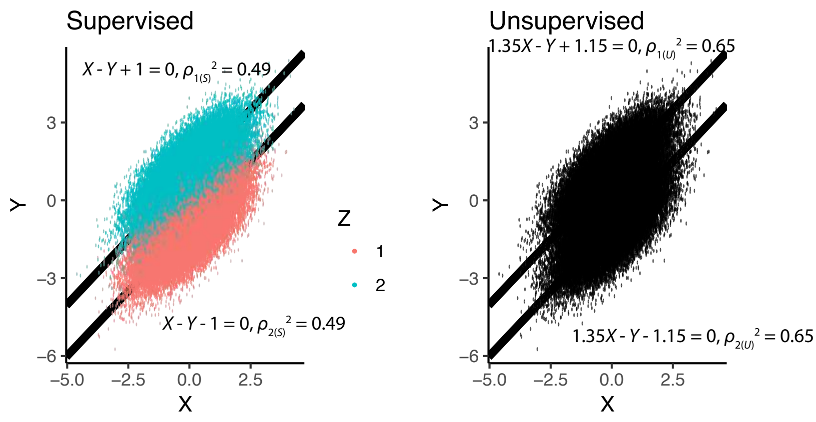
The unsupervised population line centers $B_{K(\mathcal{U})} = \{\boldsymbol{\beta}_{1(\mathcal{U})}, \ldots, \boldsymbol{\beta}_{K(\mathcal{U})}\}$

$$B_{K(\mathcal{U})} \in \arg\min_{B_K} \mathbb{E}\left[\min_{\boldsymbol{\beta}\in B_K} d_\perp^2\left((X, Y)^\mathsf{T}, \boldsymbol{\beta}\right)\right]$$

$\boldsymbol{\beta}_{k(\mathcal{U})} = \left(a_{k(\mathcal{U})}, b_{k(\mathcal{U})}, c_{k(\mathcal{U})}\right)^\mathsf{T}$: k-th unsupervised population line center

Remark: $B_{K(\mathcal{U})}$ is not unique in general

$B_{K(\mathcal{U})} \neq B_{K(\mathcal{S})}$



Random Surrogate Index $\widetilde{Z} \in \{1, \ldots, K\}$

Given the joint distribution of (X, Y)

Definition: $\widetilde{Z}$

Suppose

- unique $B_{K(\mathcal{U})} = \{\boldsymbol{\beta}_{1(\mathcal{U})}, \ldots, \boldsymbol{\beta}_{K(\mathcal{U})}\}$
- zero probability that (X, Y) is equally close to more than one $\boldsymbol{\beta}_{k(\mathcal{U})}$

We define a random surrogate index $\widetilde{Z}$ as

$$\widetilde{Z} := \underset{k \in \{1, \ldots, K\}}{\arg\min} \; d_{\perp}\left((X, Y)^{\mathsf{T}}, \boldsymbol{\beta}_{k(\mathcal{U})}\right)$$

which is uniquely determined by (X, Y) except in a measure zero set

If $d_{\perp}\left((X, Y)^{\mathsf{T}}, \boldsymbol{\beta}_{k(\mathcal{U})}\right) < \min_{r \neq k} d_{\perp}\left((X, Y)^{\mathsf{T}}, \boldsymbol{\beta}_{r(\mathcal{U})}\right)$, then $\widetilde{Z} = k$

Unsupervised Population Generalized R^2: $\rho^2_{\mathcal{G}(\mathcal{U})}$

Given the joint distribution of (X, Y)

Definition: $\rho^2_{\mathcal{G}(\mathcal{U})}$

The unsupervised population R^2 is defined as

$$\rho^2_{\mathcal{G}(\mathcal{U})} := \sum_{k=1}^{K} p_{k(\mathcal{U})} \cdot \rho^2_{k(\mathcal{U})}$$

where

$$p_{k(\mathcal{U})} := \mathbb{P}(\widetilde{Z} = k)$$

$$\rho^2_{k(\mathcal{U})} := \frac{\mathrm{cov}^2(X, Y|\widetilde{Z} = k)}{\mathrm{var}(X|\widetilde{Z} = k)\,\mathrm{var}(Y|\widetilde{Z} = k)}$$

Remark: $\rho^2_{\mathcal{G}(\mathcal{U})} \geq \rho^2_{\mathcal{G}(\mathcal{S})}$

Unsupervised Sample Line Centers

Consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$

Definition: $\widehat{B}_{K(\mathcal{U})}$

The unsupervised sample line centers $\widehat{B}_{K(\mathcal{U})} = \left\{\widehat{\boldsymbol{\beta}}_{1(\mathcal{U})}, \ldots, \widehat{\boldsymbol{\beta}}_{K(\mathcal{U})}\right\}$

$$\widehat{B}_{K(\mathcal{U})} \in \arg\min_{B_K} \frac{1}{n} \sum_{i=1}^{n} \min_{\boldsymbol{\beta} \in B_K} d_{\perp}^2 \left((X_i, Y_i)^{\mathsf{T}}, \boldsymbol{\beta}\right)$$

$\widehat{\boldsymbol{\beta}}_{k(\mathcal{U})} = \left(\widehat{a}_{k(\mathcal{U})}, \widehat{b}_{k(\mathcal{U})}, \widehat{c}_{k(\mathcal{U})}\right)^{\mathsf{T}}$: k-th unsupervised sample line center

Remark: $\widehat{B}_{K(\mathcal{U})}$ is not unique in general

K-lines Clustering Algorithm

Algorithm 1 K-lines clustering algorithm

1: **input**:

Sample: $\{(X_i, Y_i)\}_{i=1}^n$

K: number of line centers

2: **procedure** K-LINES($\{(X_i, Y_i)\}_{i=1}^n, K$)

3: Initial cluster assignment: $\mathcal{C}_1^{(0)}, \ldots, \mathcal{C}_K^{(0)}$, such that $\cup_{k=1}^K \mathcal{C}_k^{(0)} = \{1, \ldots, n\}$

4: Given the initial cluster assignment, the algorithm proceeds by alternating between two steps in each iteration. In the t-th iteration, $t = 1, 2, \ldots$

Recentering step: Calculate the cluster line centers $\widehat{\boldsymbol{\beta}}_{1(\mathcal{U})}^{(t)}, \ldots, \widehat{\boldsymbol{\beta}}_{K(\mathcal{U})}^{(t)}$ based on the cluster assignment $\mathcal{C}_1^{(t-1)}, \ldots, \mathcal{C}_K^{(t-1)}$

Assignment step: Update the cluster assignment as

$$\mathcal{C}_k^{(t)} = \left\{ i : d_\perp \left((X_i, Y_i)^\mathsf{T}, \widehat{\boldsymbol{\beta}}_{k(\mathcal{U})}^{(t)} \right) \leq d_\perp \left((X_i, Y_i)^\mathsf{T}, \widehat{\boldsymbol{\beta}}_{s(\mathcal{U})}^{(t)} \right), \forall s = 1, \ldots, K \right\}.$$

5: Stop the iteration when the cluster assignment no longer changes.

6: **output**:

Cluster assignment $\mathcal{C}_1, \ldots, \mathcal{C}_K$

K unsupervised sample line centers $\widehat{\boldsymbol{\beta}}_{1(\mathcal{U})}, \ldots, \widehat{\boldsymbol{\beta}}_{K(\mathcal{U})}$

Sample Surrogate Index $\widehat{\widetilde{Z}}_1, \ldots, \widehat{\widetilde{Z}}_n$

Consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$

Definition: $\widehat{\widetilde{Z}}_i$

Suppose

- unique $\widehat{B}_{K(\mathcal{U})} = \left\{\widehat{\beta}_{1(\mathcal{U})}, \ldots, \widehat{\beta}_{K(\mathcal{U})}\right\}$

For each (X_i, Y_i), we define its sample surrogate index

$$\widehat{\widetilde{Z}}_i := \underset{k \in \{1,\ldots,K\}}{\arg\min}\ d_{\perp}\left((X_i, Y_i)^{\mathsf{T}}, \widehat{\beta}_{k(\mathcal{U})}\right), \quad i = 1, \ldots, n$$

which is uniquely determined by the sample

$$\widehat{\widetilde{Z}}_i = k \iff i \in \mathcal{C}_k\,,$$

$\mathcal{C}_k$: the k-th cluster output by the K-lines clustering algorithm, assuming the global minimum is achieved

Unsupervised Sample Generalized R^2: $R^2_{\mathcal{G}(\mathcal{U})}$

Consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$

Definition: $R^2_{\mathcal{G}(\mathcal{U})}$

The unsupervised sample generalized R^2 is defined as

$$R^2_{\mathcal{G}(\mathcal{U})} := \sum_{k=1}^{K} \widehat{p}_{k(\mathcal{U})} \cdot \widehat{\rho}^2_{k(\mathcal{U})}$$

where

$$\widehat{p}_{k(\mathcal{U})} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)$$

$$\widehat{\rho}^2_{k(\mathcal{U})} = \frac{\left[\sum_{i=1}^n \left(X_i - \bar{X}_{k(\mathcal{U})}\right)\left(Y_i - \bar{Y}_{k(\mathcal{U})}\right) \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)\right]^2}{\left[\sum_{i=1}^n \left(X_i - \bar{X}_{k(\mathcal{U})}\right)^2 \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)\right]\left[\sum_{i=1}^n \left(Y_i - \bar{Y}_{k(\mathcal{U})}\right)^2 \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)\right]}$$

with

- $\bar{X}_{k(\mathcal{U})} = \frac{1}{n_{k(\mathcal{U})}} \sum_{i=1}^n X_i \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)$; $\bar{Y}_{k(\mathcal{U})} = \frac{1}{n_{k(\mathcal{U})}} \sum_{i=1}^n Y_i \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)$
- $n_{k(\mathcal{U})} = \sum_{i=1}^n \mathbb{I}\left(\widehat{\widehat{Z}}_i = k\right)$

Choose K in the Unsupervised Scenario

Criteria

1. Average within-cluster sum of perpendicular distances

Definition: $W(B_K, P_n)$

$$W(B_K, P_n) := \frac{1}{n}\sum_{i=1}^{n} \min_{\boldsymbol{\beta}\in B_K} d_\perp^2\left((X_i, Y_i)^\mathsf{T}, \boldsymbol{\beta}\right)$$
$$= \int \min_{\boldsymbol{\beta}\in B_K} d_\perp^2\left((x, y)^\mathsf{T}, \boldsymbol{\beta}\right) P_n\left((dx, dy)^\mathsf{T}\right) ,$$

P_n: the empirical measure by placing mass n^{-1} at each (X_i, Y_i)

Criteria

2. Akaike information criterion (AIC)

Definition: AIC(K)

$$\mathrm{AIC}(K) := 12K - 2\sum_{i=1}^{n} \log p\left(X_i, Y_i \;\middle|\; \left\{\widehat{p}_{k(\mathcal{U})}, \widehat{\boldsymbol{\mu}}_{k(\mathcal{U})}, \widehat{\boldsymbol{\Sigma}}_{k(\mathcal{U})}\right\}_{k=1}^{K}\right)$$

where

$$p\left(X_i, Y_i \;\middle|\; \left\{\widehat{p}_{k(\mathcal{U})}, \widehat{\boldsymbol{\mu}}_{k(\mathcal{U})}, \widehat{\boldsymbol{\Sigma}}_{k(\mathcal{U})}\right\}_{k=1}^{K}\right)$$
$$= \sum_{k=1}^{K} \widehat{p}_{k(\mathcal{U})} \frac{\exp\left\{-\frac{1}{2}\left((X_i, Y_i)^{\mathsf{T}} - \widehat{\boldsymbol{\mu}}_{k(\mathcal{U})}\right)^{\mathsf{T}} \widehat{\boldsymbol{\Sigma}}_{k(\mathcal{U})}^{-1} \left((X_i, Y_i)^{\mathsf{T}} - \widehat{\boldsymbol{\mu}}_{k(\mathcal{U})}\right)\right\}}{2\pi\sqrt{\left|\widehat{\boldsymbol{\Sigma}}_{k(\mathcal{U})}\right|}}$$

Asymptotic Distribution of $\rho^2_{\mathcal{G}(\mathcal{S})}$ — General

Define

$$\mu_{X^cY^d,k(\mathcal{S})} = \mathbb{E}\left[\left(\frac{X-\mathbb{E}[X|Z=k]}{\sqrt{\mathrm{var}(X|Z=k)}}\right)^c\left(\frac{Y-\mathbb{E}[Y|Z=k]}{\sqrt{\mathrm{var}(Y|Z=k)}}\right)^d \middle| Z=k\right], \quad c,d\in\mathbb{N}$$

Theorem:

Assume $\mu_{X^4,k(\mathcal{S})} < \infty$ and $\mu_{Y^4,k(\mathcal{S})} < \infty$ for all $k = 1,\ldots,K$. Then

$$\sqrt{n}\left(R^2_{\mathcal{G}(\mathcal{S})} - \rho^2_{\mathcal{G}(\mathcal{S})}\right) \xrightarrow{d} \mathcal{N}\left(0, \gamma^2_{(\mathcal{S})}\right)$$

where

$$\gamma^2_{(\mathcal{S})} = \sum_{k=1}^{K}\left(A_{k(\mathcal{S})} + B_{k(\mathcal{S})}\right) + 2\sum\sum_{1\le k<r\le K} C_{kr(\mathcal{S})}$$

$$A_{k(\mathcal{S})} = p_{k(\mathcal{S})}\Big[\rho^4_{k(\mathcal{S})}\left(\mu_{X^4,k(\mathcal{S})} + 2\mu_{X^2Y^2,k(\mathcal{S})} + \mu_{Y^4,k(\mathcal{S})}\right) - 4\rho^3_{k(\mathcal{S})}\left(\mu_{X^3Y,k(\mathcal{S})} + \mu_{XY^3,k(\mathcal{S})}\right)$$
$$+4\rho^2_{k(\mathcal{S})}\mu_{X^2Y^2,k(\mathcal{S})}\Big]$$

$$B_{k(\mathcal{S})} = p_{k(\mathcal{S})}\left(1 - p_{k(\mathcal{S})}\right)\rho^4_{k(\mathcal{S})}$$

$$C_{kr(\mathcal{S})} = -\,p_{k(\mathcal{S})}\,p_{r(\mathcal{S})}\,\rho^2_{k(\mathcal{S})}\,\rho^2_{r(\mathcal{S})}$$

Asymptotic Distribution of $\rho^2_{\mathcal{G}(\mathcal{S})}$ — Bivariate Gaussian Mixture

Corollary:

In the special case where $(X, Y)|(Z = k)$ follows a bivariate Gaussian distribution for all $k = 1, \ldots, K$, $\gamma^2_{(\mathcal{S})}$ becomes

$$\gamma^2_{(\mathcal{S})} = \sum_{k=1}^{K} \left[4\, p_{k(\mathcal{S})}\, \rho^2_{k(\mathcal{S})} \left(1 - \rho^2_{k(\mathcal{S})}\right)^2 + p_{k(\mathcal{S})} \left(1 - p_{k(\mathcal{S})}\right) \rho^4_{k(\mathcal{S})} \right] - 2 \sum\sum_{1 \le k < r \le K} p_{k(\mathcal{S})}\, p_{r(\mathcal{S})}\, \rho^2_{k(\mathcal{S})}\, \rho^2_{r(\mathcal{S})}$$

which only depends on $p_{k(\mathcal{S})}$ and $\rho^2_{k(\mathcal{S})}$, $k = 1, \ldots, K$

Strong Consistency of the K-lines Clustering

Theorem:

Suppose

- $\int \left\|(x, y)^\mathsf{T}\right\|^2 P\left((dx, dy)^\mathsf{T}\right) < \infty$
- for each $k = 1, \ldots, K$, there is unique $B_{k(\mathcal{U})} = \arg\min_{B_k} W(B_k, P)$

As the sample size $n \to \infty$,

$$\widehat{B}_{K(\mathcal{U})} \to B_{K(\mathcal{U})} \text{ almost surely}$$

and

$$W(\widehat{B}_{K(\mathcal{U})}, P_n) \to W(B_{K(\mathcal{U})}, P) \text{ almost surely}$$

Asymptotic Distribution of $\rho^2_{\mathcal{G}(\mathcal{U})}$ — General

Define

$$\mu_{X^cY^d,k(\mathcal{U})} = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X|\widetilde{Z}=k]}{\sqrt{\mathrm{var}(X|\widetilde{Z}=k)}}\right)^c \left(\frac{Y - \mathbb{E}[Y|\widetilde{Z}=k]}{\sqrt{\mathrm{var}(Y|\widetilde{Z}=k)}}\right)^d \middle| \widetilde{Z}=k\right], \quad c,d \in \mathbb{N}$$

Theorem:

Assume $\mu_{X^4,k(\mathcal{U})} < \infty$ and $\mu_{Y^4,k(\mathcal{U})} < \infty$ for all $k = 1, \dots, K$. Then

$$\sqrt{n}\left(R^2_{\mathcal{G}(\mathcal{U})} - \rho^2_{\mathcal{G}(\mathcal{U})}\right) \xrightarrow{d} \mathcal{N}\left(0, \gamma^2_{(\mathcal{U})}\right)$$

where

$$\gamma^2_{(\mathcal{U})} = \sum_{k=1}^{K}\left(A_{k(\mathcal{U})} + B_{k(\mathcal{U})}\right) + 2\sum\sum_{1\le k<r\le K} C_{kr(\mathcal{U})}$$

$$A_{k(\mathcal{U})} = p_{k(\mathcal{U})}\Big[\rho^4_{k(\mathcal{U})}\left(\mu_{X^4,k(\mathcal{U})} + 2\mu_{X^2Y^2,k(\mathcal{U})} + \mu_{Y^4,k(\mathcal{U})}\right) - 4\rho^3_{k(\mathcal{U})}\left(\mu_{X^3Y,k(\mathcal{U})} + \mu_{XY^3,k(\mathcal{U})}\right)$$
$$+4\rho^2_{k(\mathcal{U})}\mu_{X^2Y^2,k(\mathcal{U})}\Big]$$

$$B_{k(\mathcal{U})} = p_{k(\mathcal{U})}\left(1 - p_{k(\mathcal{U})}\right)\rho^4_{k(\mathcal{U})}$$

$$C_{kr(\mathcal{U})} = -\,p_{k(\mathcal{U})}\,p_{r(\mathcal{U})}\,\rho^2_{k(\mathcal{U})}\,\rho^2_{r(\mathcal{U})}$$

Asymptotic Distribution of $\rho^2_{\mathcal{G}(\mathcal{U})}$ — Bivariate Gaussian Mixture

Corollary:

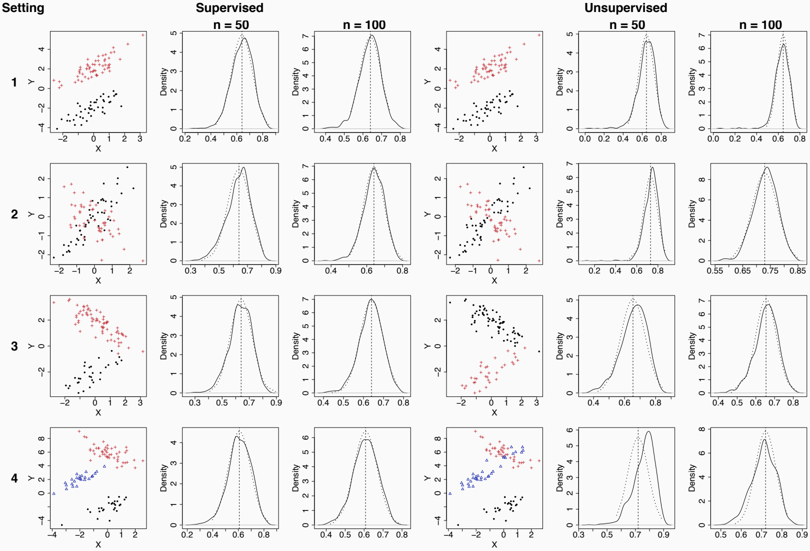
In the special case where $(X, Y)|(\widetilde{Z} = k)$ follows a bivariate Gaussian distribution for all $k = 1, \ldots, K$, $\gamma^2_{(\mathcal{U})}$ becomes

$$\gamma^2_{(\mathcal{U})} = \sum_{k=1}^{K} \left[4\, p_{k(\mathcal{U})}\, \rho^2_{k(\mathcal{U})} \left(1 - \rho^2_{k(\mathcal{U})}\right)^2 + p_{k(\mathcal{U})} \left(1 - p_{k(\mathcal{U})}\right) \rho^4_{k(\mathcal{U})} \right] - 2 \sum\sum_{1 \le k < r \le K} p_{k(\mathcal{U})}\, p_{r(\mathcal{U})}\, \rho^2_{k(\mathcal{U})}\, \rho^2_{r(\mathcal{U})}$$

which only depends on $p_{k(\mathcal{U})}$ and $\rho^2_{k(\mathcal{U})}$, $k = 1, \ldots, K$

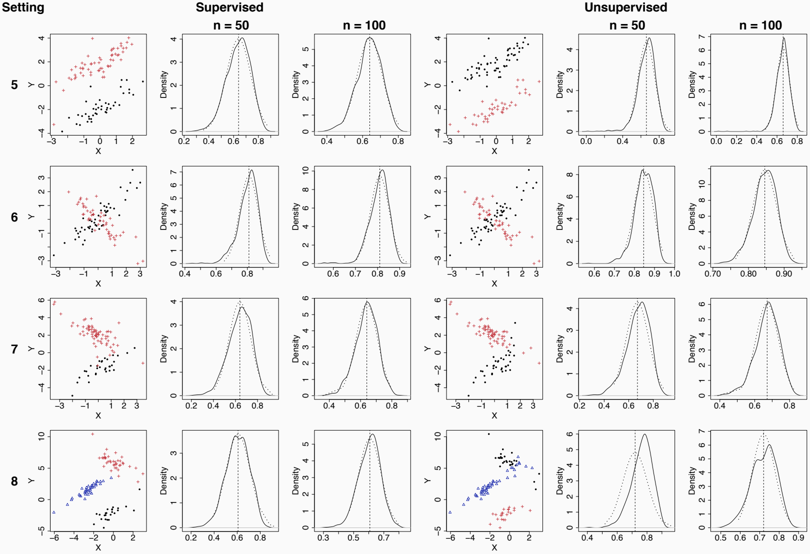
Simulation: Numerical Verification of Asymptotic Distributions

$$(X, Y)|(Z = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



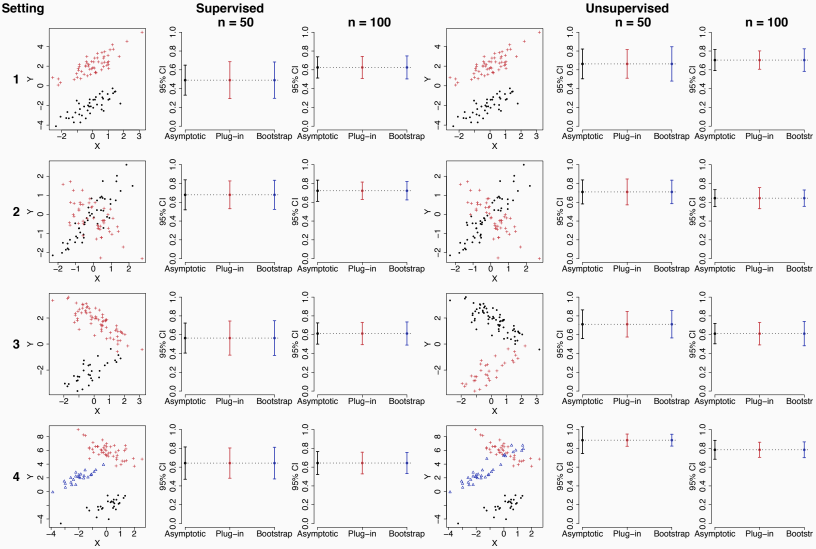
Simulation: Numerical Verification of Asymptotic Distributions

$$(X, Y)|(Z = k) \sim t_{\nu_k}(\mu_k, \Sigma_k)$$



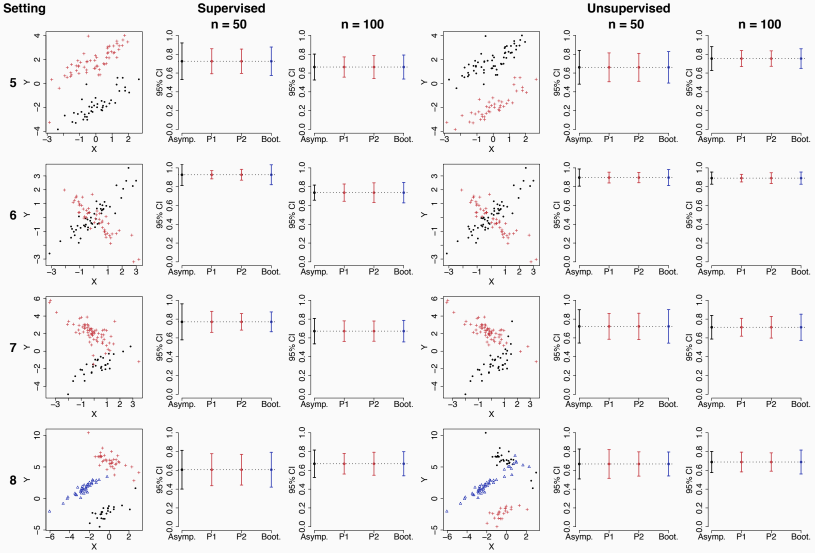
Simulation: Numerical Verification of Confidence Intervals

$$(X, Y)|(Z = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



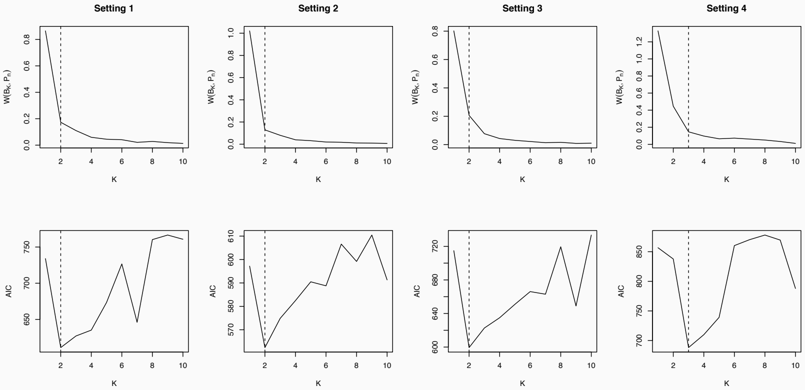
Simulation: Numerical Verification of Confidence Intervals

$$(X, Y)|(Z = k) \sim t_{\nu_k}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



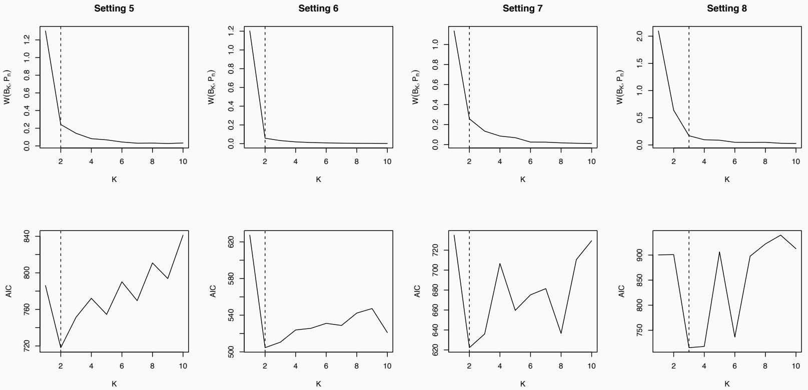
Simulation: Choose K

$$(X, Y)|(Z = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

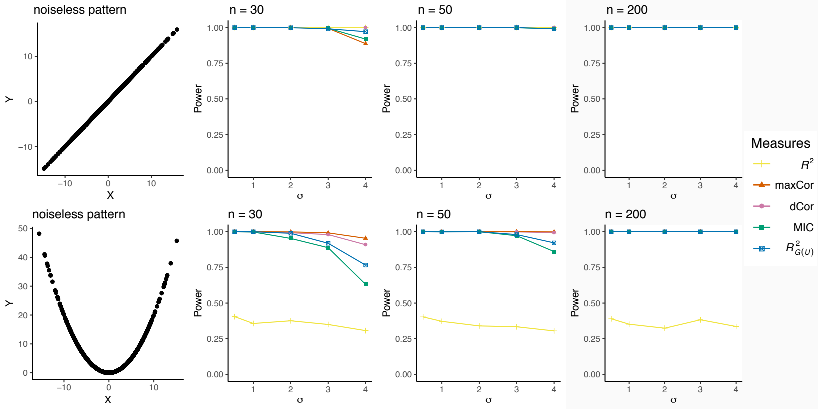


Simulation: Choose K

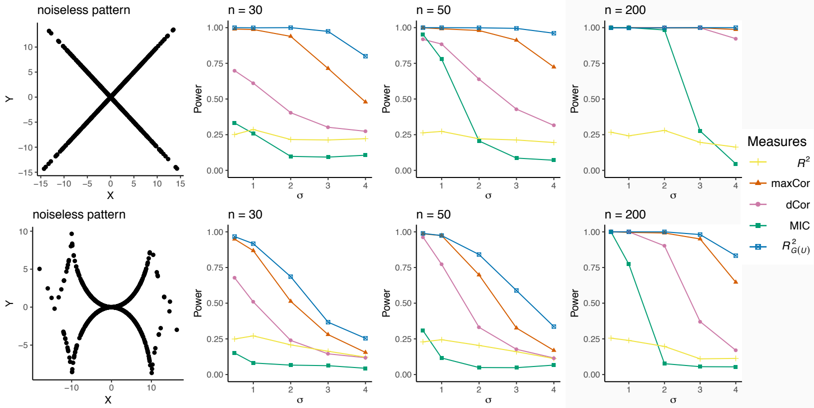
$$(X, Y)|(Z = k) \sim t_{\nu_k}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



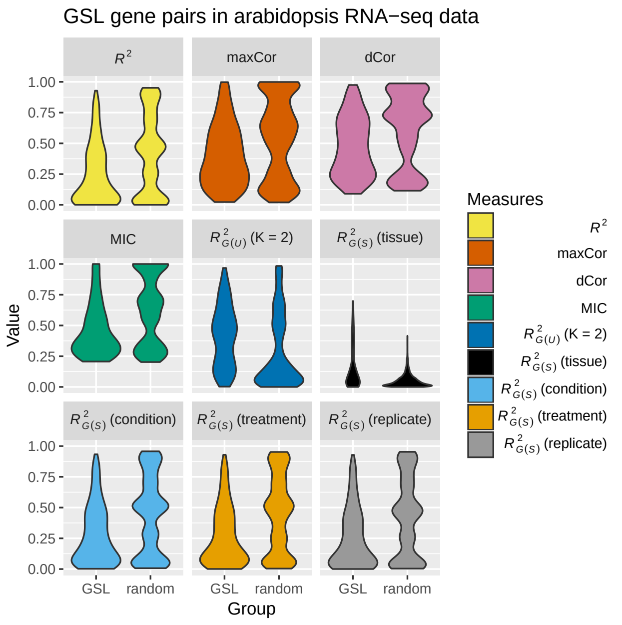
Simulation: Power Analysis



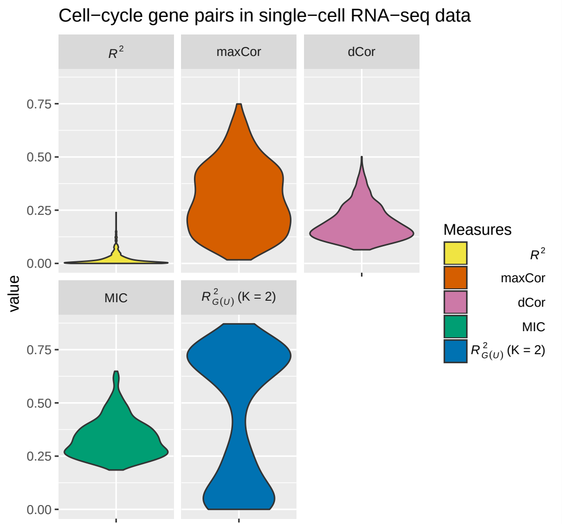
Simulation: Power Analysis



Real Data Application 1



Real Data Application 2



- *Cdc25b*-*Lats2* receive the highest $R^2_{\mathcal{G}(\mathcal{U})}$ value (Mukai et al., 2015)
- *Lats2* appears in the top 25% pairs that have the highest $R^2_{\mathcal{G}(\mathcal{U})}$ values (Yabuta et al., 2007)

Summary & Future Directions

Summary

- A mixture of linear dependences
- Generalized (population and sample) R^2 measures
  - Supervised scenario
  - Unsupervised scenario
- Statistical inference of the generalized population R^2 measures
- K-lines algorithm

Future Directions

- A sequential test for $K = 1, 2, \ldots, K_{\text{max}}$
- Rank-based generalized R^2 measures

Paper and Software

Generalized R^2 Measures for a Mixture of Bivariate Linear Dependences

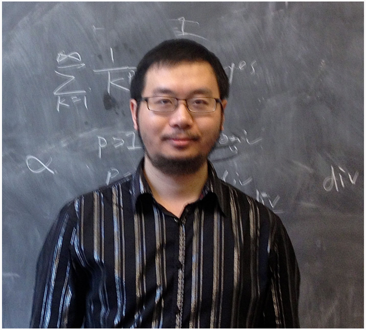
by Jingyi Jessica Li, Xin Tong, and Peter J. Bickel

arXiv:1811.09965

R package `gR2`

`https://github.com/lijy03/gR2`

Acknowledgements



Xin Tong
(USC)

Peter J. Bickel
(UC Berkeley)

NIH NIGMS

NSF

Alfred P. Sloan FOUNDATION

Johnson & Johnson WiSTEM2D
Women in Science, Technology, Engineering, Math, Manufacturing and Design

PhRMA
RESEARCH • PROGRESS • HOPE