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

Detecting Novel Associations in Large Data Sets

David N. Reshef\textsuperscript{1,2,3,*,†}, Yakir A. Reshef\textsuperscript{2,4,*,†}, Hilary K. Finucane\textsuperscript{5}, Sharon R. Grossman\textsuperscript{2,6}, Gilean McVean\textsuperscript{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

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

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

These maximal correlation values < 1 were due to lack of convergence.
Motivation: Maximal Information Coefficient

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Maximal Information Coefficient (MIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Lines</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Line and Parabola</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>X</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Ellipse</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Sinusoid (Mixture of two signals)</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Non-coexistence</td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

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) ≈ 0
Motivation: Simpson’s Paradox

Pearson cor (red) ≈ 0.8
Pearson cor (blue) ≈ −0.8
Pearson cor (all) ≈ 0
# Review: Scalar-valued Association Measures

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

<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functional</strong> (1-to-1)</td>
<td><strong>Linear</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Monotone</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>General</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dependent</strong></td>
<td>Hoeffding’s $D$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Review: Scalar-valued Association Measures

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

<table>
<thead>
<tr>
<th>Measures for Relationship Type</th>
<th>Interpretability</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional (1-to-1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>best</td>
<td>worst</td>
</tr>
<tr>
<td>Monotone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>worst</td>
<td>best</td>
</tr>
<tr>
<td>Dependent</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Mixture of linear dependences
- Widespread
- Easy to interpret
- Calling for a new powerful measure
Review: Scalar-valued Association Measures

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

<table>
<thead>
<tr>
<th>Measures for Relationship Type</th>
<th>Interpretability</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional (1-to-1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>best</td>
<td>worst</td>
</tr>
<tr>
<td>Monotone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>worst</td>
<td>best</td>
</tr>
</tbody>
</table>

Mixture of linear dependences

- Widespread
- Easy to interpret
- Calling for a new powerful measure
Review: Mixture of Linear Dependences

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

\[
p_k(S) := \mathbb{P}(Z = k), \quad k = 1, \ldots, K, \text{ with } \sum_{k=1}^{K} p_k(S) = 1.
\]

and

\[
\rho_k(S) := \frac{\text{cov}(X, Y|Z = k)}{\sqrt{\text{var}(X|Z = k)}} \frac{1}{\sqrt{\text{var}(Y|Z = k)}}
\]

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

**Definition: \(\rho^2_G(S)\)**

The supervised population generalized \(R^2\) is defined as

\[
\rho^2_G(S) := \mathbb{E}_Z \left[ \rho^2_X(S) \right] = \mathbb{E}_Z \left[ \frac{\text{cov}^2(X, Y|Z)}{\text{var}(X|Z)\text{var}(Y|Z)} \right] = \sum_{k=1}^{K} p_k(S) \cdot \rho^2_k(S)
\]
K-line Interpretation the Supervised Scenario

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

$$\{(x, y)^T : ax + by + c = 0, \text{ where } a, b, c \in \mathbb{R} \text{ with } a \neq 0 \text{ or } b \neq 0\} \subset \mathbb{R}^2$$
**K-line Interpretation the Supervised Scenario**

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

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

- **Perpendicular distance** between $(x, y)^T$ and $\beta$ is

$$d_{\perp} : \mathbb{R}^2 \times \mathbb{R}^3 \mapsto \mathbb{R} :$$

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

*Symmetric* between $x$ and $y$
- Denote by $\beta = (a, b, c)^T$ a line

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

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

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

**Symmetric** between $x$ and $y$

**Definition: Supervised Population $k$-th Line Center**

$$\beta_{k(S)} = \arg \min_{\beta} \mathbb{E} \left[ d_\bot^2 ((X, Y)^T, \beta) \mid Z = k \right]$$
**$K$-line Interpretation the Supervised Scenario**

**Definition: Supervised Population $k$-th Line Center**

\[
\beta_{k(S)} = \arg \min_{\beta} \mathbb{E} \left[ d_{\perp}^2 ((X, Y)^T, \beta) \left| Z = k \right. \right]
\]

corresponds to the **first principal component** of

\[
\Sigma_{k(S)} := \begin{bmatrix}
\text{var}(X|Z = k) & \text{cov}(X, Y|Z = k) \\
\text{cov}(X, Y|Z = k) & \text{var}(Y|Z = k)
\end{bmatrix}
\]

(Jolliffe, 2011)

\[
B_{K(S)} = \{\beta_1(S), \ldots, \beta_K(S)\}: \text{ supervised population line centers}
\]
Consider a sample \((X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\)

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

The supervised sample generalized \(R^2\) is defined as

\[
R^2_{G(S)} := \sum_{k=1}^{K} \hat{p}_k(S) \cdot \hat{\rho}^2_k(S)
\]

where

\[
\hat{p}_k(S) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Z_i = k)
\]

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

with

- \(\bar{X}_k(S) = \frac{1}{n_k(S)} \sum_{i=1}^{n} X_i \mathbb{I}(Z_i = k)\);
- \(\bar{Y}_k(S) = \frac{1}{n_k(S)} \sum_{i=1}^{n} Y_i \mathbb{I}(Z_i = k)\)
- \(n_k(S) = \sum_{i=1}^{n} \mathbb{I}(Z_i = k)\)
Unsupervised Population Line Centers

Given the joint distribution of \((X, Y)\)

**Definition:** \(B_K(\mathcal{U})\)

The **unsupervised population line centers** \(B_K(\mathcal{U}) = \{\beta_1(\mathcal{U}), \ldots, \beta_K(\mathcal{U})\}\)

\[
B_K(\mathcal{U}) \in \arg \min \mathbb{E} \left[ \min_{\beta \in B_K} d^2_\perp ((X, Y)^T, \beta) \right]
\]

\(\beta_k(\mathcal{U}) = (a_k(\mathcal{U}), b_k(\mathcal{U}), c_k(\mathcal{U}))^T: k\)-th unsupervised population line center

**Remark:** \(B_K(\mathcal{U})\) is not unique in general
$B_K(U) \neq B_K(S)$

**Supervised**

$X - Y + 1 = 0, \rho_{1(S)}^2 = 0.49$

$X - Y - 1 = 0, \rho_{2(S)}^2 = 0.49$

**Unsupervised**

$1.35X - Y + 1.15 = 0, \rho_{1(U)}^2 = 0.65$

$1.35X - Y - 1.15 = 0, \rho_{2(U)}^2 = 0.65$
Random Surrogate Index $\tilde{Z} \in \{1, \ldots, K\}$

Given the joint distribution of $(X, Y)$

**Definition: $\tilde{Z}$**

Suppose

- unique $B_K(U) = \{\beta_1(U), \ldots, \beta_K(U)\}$
- zero probability that $(X, Y)$ is equally close to more than one $\beta_k(U)$

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

$$\tilde{Z} := \arg \min_{k \in \{1, \ldots, K\}} d_{\perp} ((X, Y)^T, \beta_k(U))$$

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

If $d_{\perp} ((X, Y)^T, \beta_k(U)) < \min_{r \neq k} d_{\perp} ((X, Y)^T, \beta_r(U))$, then $\tilde{Z} = k$
Given the joint distribution of \((X, Y)\)

**Definition:** \(\rho_{G(U)}^2\)

The unsupervised population \(R^2\) is defined as

\[
\rho_{G(U)}^2 := \sum_{k=1}^{K} p_k(U) \cdot \rho_k(U)
\]

where

\[
p_k(U) := \mathbb{P}(\tilde{Z} = k)
\]

\[
\rho_k(U) := \frac{\text{cov}^2(X, Y|\tilde{Z} = k)}{\text{var}(X|\tilde{Z} = k) \cdot \text{var}(Y|\tilde{Z} = k)}
\]

**Remark:** \(\rho_{G(U)}^2 \geq \rho_{G(S)}^2\)
Consider a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\)

**Definition: \(\hat{\mathcal{B}}_K(\mathcal{U})\)**

The unsupervised sample line centers \(\hat{\mathcal{B}}_K(\mathcal{U}) = \{\hat{\beta}_1(\mathcal{U}), \ldots, \hat{\beta}_K(\mathcal{U})\}\)

\[
\hat{\mathcal{B}}_K(\mathcal{U}) \in \arg \min_{\mathcal{B}_K} \frac{1}{n} \sum_{i=1}^{n} \min_{\beta \in \mathcal{B}_K} d_\perp^2 ((X_i, Y_i)^T, \beta)
\]

\[
\hat{\beta}_k(\mathcal{U}) = \left(\hat{a}_k(\mathcal{U}), \hat{b}_k(\mathcal{U}), \hat{c}_k(\mathcal{U})\right)^T : k\text{-th unsupervised sample line center}
\]

**Remark:** \(\hat{\mathcal{B}}_K(\mathcal{U})\) is not unique in general
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: $C_1^{(0)}, \ldots, C_K^{(0)}$, such that $\bigcup_{k=1}^K C_k^{(0)} = \{1, \ldots, n\}$

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 $\hat{\beta}_1^{(t)}(U), \ldots, \hat{\beta}_K^{(t)}(U)$ based on the cluster assignment $C_1^{(t-1)}, \ldots, C_K^{(t-1)}$

   **Assignment step:** Update the cluster assignment as

   $$C_k^{(t)} = \left\{ i : d_\perp \left( (X_i, Y_i)^T, \hat{\beta}_k^{(t)}(U) \right) \leq d_\perp \left( (X_i, Y_i)^T, \hat{\beta}_s^{(t)}(U) \right), \forall s = 1, \ldots, K \right\}.$$

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

6: output:
   Cluster assignment $C_1, \ldots, C_K$
   $K$ unsupervised sample line centers $\hat{\beta}_1(U), \ldots, \hat{\beta}_K(U)$
Consider a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\)

**Definition: \(\hat{Z}_i\)**

Suppose

- unique \(\hat{B}_{K(U)} = \{\hat{\beta}_1(U), \ldots, \hat{\beta}_K(U)\}\)

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

\[
\hat{Z}_i := \arg \min_{k \in \{1, \ldots, K\}} d_\perp \left((X_i, Y_i)^T, \hat{\beta}_k(U)\right), \ i = 1, \ldots, n
\]

which is uniquely determined by the sample

\[
\hat{Z}_i = k \iff i \in C_k,
\]

\(C_k\): the \(k\)-th cluster output by the \(K\)-lines clustering algorithm, assuming the global minimum is achieved
Unsupervised Sample Generalized $R^2$: $R^2_G(U)$

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

**Definition: $R^2_G(U)$**

The unsupervised sample generalized $R^2$ is defined as

$$R^2_G(U) := \sum_{k=1}^{K} \hat{p}_k(U) \cdot \hat{\rho}^2_k(U)$$

where

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

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

with

- $\bar{X}_k(U) = \frac{1}{n_k(U)} \sum_{i=1}^{n} X_i \mathbb{I} \left( \hat{Z}_i = k \right)$; $\bar{Y}_k(U) = \frac{1}{n_k(U)} \sum_{i=1}^{n} Y_i \mathbb{I} \left( \hat{Z}_i = k \right)$
- $n_k(U) = \sum_{i=1}^{n} \mathbb{I} \left( \hat{Z}_i = k \right)$
Choose $K$ in the Unsupervised Scenario

Criteria

1. Average within-cluster sum of perpendicular distances

**Definition:** $W(B_K, P_n)$

$$W(B_K, P_n) := \frac{1}{n} \sum_{i=1}^{n} \min_{\beta \in B_K} d^2_\perp \left((X_i, Y_i)^T, \beta\right)$$

$$= \int \min_{\beta \in B_K} d^2_\perp \left((x, y)^T, \beta\right) P_n \left((dx, dy)^T\right) ,$$

$P_n$: the empirical measure by placing mass $n^{-1}$ at each $(X_i, Y_i)$
Choose $K$ in the Unsupervised Scenario

Criteria

2. Akaike information criterion (AIC)

**Definition: AIC($K$)**

\[
\text{AIC}(K) := 12K - 2 \sum_{i=1}^{n} \log p \left( X_i, Y_i \bigg| \left\{ \hat{p}_k(U), \hat{\mu}_k(U), \hat{\Sigma}_k(U) \right\}_{k=1}^{K} \right)
\]

where

\[
p \left( X_i, Y_i \bigg| \left\{ \hat{p}_k(U), \hat{\mu}_k(U), \hat{\Sigma}_k(U) \right\}_{k=1}^{K} \right) = \sum_{k=1}^{K} \hat{p}_k(U) \exp \left\{ -\frac{1}{2} \left( (X_i, Y_i)^T - \hat{\mu}_k(U) \right)^T \hat{\Sigma}_k(U)^{-1} \left( (X_i, Y_i)^T - \hat{\mu}_k(U) \right) \right\}
\]

\[
= 2\pi \sqrt{\left| \hat{\Sigma}_k(U) \right|}
\]
Asymptotic Distribution of $\rho^2_G(S)$ — General

Define

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

**Theorem:**

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

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

where

$$\gamma^2(S) = \sum_{k=1}^K (A_k(S) + B_k(S)) + 2 \sum_{1 \leq k < r \leq K} C_{kr}(S)$$

$$A_k(S) = p_k(S) \left[ \rho^4_k(S) \left( \mu_{X^4,k(S)} + 2\mu_XY^2,k(S) + \mu_{Y^4,k(S)} \right) - 4\rho^3_k(S) \left( \mu_XY,k(S) + \mu_XY^3,k(S) \right) + 4\rho^2_k(S)\mu_XY^2,k(S) \right]$$

$$B_k(S) = p_k(S) (1 - p_k(S)) \rho^4_k(S)$$

$$C_{kr}(S) = -p_k(S) p_r(S) \rho^2_k(S) \rho^2_r(S)$$
Corollary:

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

\[
\gamma^2_{(S)} = \sum_{k=1}^{K} \left[ 4 p_k(S) \rho^2_k(S) \left(1 - \rho^2_k(S)\right)^2 + p_k(S) \left(1 - p_k(S)\right) \rho^4_k(S) \right] \\
- 2 \sum_{1 \leq k < r \leq K} p_k(S) p_r(S) \rho^2_k(S) \rho^2_r(S)
\]

which only depends on \(p_k(S)\) and \(\rho^2_k(S)\), \(k = 1, \ldots, K\)
Theorem:

Suppose

1. \( \int \| (x, y)^T \|^2 P((dx, dy)^T) < \infty \)
2. 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 \( n \to \infty \),

\[ \hat{B}_K(\mathcal{U}) \to B_K(\mathcal{U}) \text{ almost surely} \]

and

\[ W(\hat{B}_K(\mathcal{U}), P_n) \to W(B_K(\mathcal{U}), P) \text{ almost surely} \]
Asymptotic Distribution of $\rho^2_G(U)$ — General

Define

$$\mu_{X^c Y^d, k(U)} = \mathbb{E} \left[ \left( \frac{X - \mathbb{E}[X|\tilde{Z} = k]}{\sqrt{\text{var}(X|\tilde{Z} = k)}} \right)^c \left( \frac{Y - \mathbb{E}[Y|\tilde{Z} = k]}{\sqrt{\text{var}(Y|\tilde{Z} = k)}} \right)^d \right| \tilde{Z} = k \right], \ c, d \in \mathbb{N}$$

Theorem:

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

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

where

$$\gamma^2(U) = \sum_{k=1}^{K} (A_k(U) + B_k(U)) + 2 \sum_{1 \leq k < r \leq K} C_{kr}(U)$$

$$A_k(U) = p_k(U) \left[ \rho^4_k(U) \left( \mu_{X^4, k(U)} + 2\mu_{X^2 Y^2, k(U)} + \mu_{Y^4, k(U)} \right) - 4\rho^3_k(U) \left( \mu_{X^3 Y, k(U)} + \mu_{X Y^3, k(U)} \right) + 4\rho^2_k(U) \mu_{X^2 Y^2, k(U)} \right]$$

$$B_k(U) = p_k(U) \left( 1 - p_k(U) \right) \rho^4_k(U)$$

$$C_{kr}(U) = - p_k(U) p_r(U) \rho^2_k(U) \rho^2_r(U)$$

27
Asymptotic Distribution of $\rho^2_{G(U)}$ — Bivariate Gaussian Mixture

**Corollary:**

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

$$\gamma^2_{(U)} = \sum_{k=1}^{K} \left[ 4 p_k(U) \rho^2_k(U) \left( 1 - \rho^2_k(U) \right)^2 + p_k(U) \left( 1 - p_k(U) \right) \rho^4_k(U) \right]$$

$$- 2 \sum_{1 \leq k < r \leq K} p_k(U) p_r(U) \rho^2_k(U) \rho^2_r(U)$$

which only depends on $p_k(U)$ and $\rho^2_k(U), k = 1, \ldots, K$.
Simulation: Numerical Verification of Asymptotic Distributions

\[(X, Y) | (Z = k) \sim \mathcal{N}(\mu_k, \Sigma_k)\]
Simulation: Numerical Verification of Asymptotic Distributions

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

<table>
<thead>
<tr>
<th>Setting</th>
<th>Supervised</th>
<th>Unsupervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 50</td>
<td>n = 100</td>
</tr>
<tr>
<td>5</td>
<td><img src="image" alt="Supervised Setting 5" /></td>
<td><img src="image" alt="Unsupervised Setting 5" /></td>
</tr>
<tr>
<td>6</td>
<td><img src="image" alt="Supervised Setting 6" /></td>
<td><img src="image" alt="Unsupervised Setting 6" /></td>
</tr>
<tr>
<td>7</td>
<td><img src="image" alt="Supervised Setting 7" /></td>
<td><img src="image" alt="Unsupervised Setting 7" /></td>
</tr>
<tr>
<td>8</td>
<td><img src="image" alt="Supervised Setting 8" /></td>
<td><img src="image" alt="Unsupervised Setting 8" /></td>
</tr>
</tbody>
</table>
Simulation: Numerical Verification of Confidence Intervals

\[(X, Y) | (Z = k) \sim \mathcal{N}(\mu_k, \Sigma_k)\]
Simulation: Numerical Verification of Confidence Intervals

\[(X, Y)|(Z = k) \sim t_{\nu_k} (\mu_k, \Sigma_k)\]
Simulation: Choose $K$

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

Setting 1

Setting 2

Setting 3

Setting 4
Simulation: Choose $K$

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

Noiseless pattern

Y

noiseless pattern

Y

n = 30

n = 50

n = 200

Power

Power

Power

Measures

$R^2$

maxCor

dCor

MIC

$R^2_{G(u)}$
Simulation: Power Analysis

![Graphs showing power analysis for different noiseless patterns and sample sizes](image)
Real Data Application 1

GSL gene pairs in Arabidopsis RNA-seq data

<table>
<thead>
<tr>
<th>Group</th>
<th>Value</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSL</td>
<td></td>
<td>$R^2$ (condition)</td>
</tr>
<tr>
<td>random</td>
<td></td>
<td>$R^2_G(U)$ ($K = 2$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R^2_G(S)$ (tissue)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R^2_G(S)$ (condition)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R^2_G(S)$ (treatment)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R^2_G(S)$ (replicate)</td>
</tr>
</tbody>
</table>

- $R^2$: Coefficient of determination
- maxCor: Maximum correlation
- dCor: Distance correlation
- MIC: Mutual information coefficient

GSL gene pairs in Arabidopsis RNA-seq data.
Cell-cycle gene pairs in single-cell RNA-seq data

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