

A new coefficient of correlation

Sourav Chatterjee

木村 大地

- NTT Communications

Paper Information

- [\[PDF\] A New Coefficient of Correlation | Semantic Scholar](#)
- Author: Sourav Chatterjee (Professor of Mathematics and Statistics, Stanford University)
- Publication: Journal of the American Statistical Association, 116.536 (2021)



[Sourav Chatterjee - Wikipedia](#)



Sourav Chatterjee

Professor of Mathematics and Statistics, [Stanford University](#).
確認したメール アドレス: stanford.edu



[自分のプロフィールを作成](#)

タイトル

引用先

年

[Spectral clustering and the high-dimensional stochastic blockmodel](#)

K Rohe, S Chatterjee, B Yu

1115

2011

[Matrix estimation by universal singular value thresholding](#)

S Chatterjee

636

2015

[Estimating and understanding exponential random graph models](#)

S Chatterjee, P Diaconis
The Annals of Statistics 41 (5), 2428-2461

451

2013

[Random graphs with a given degree sequence](#)

S Chatterjee, P Diaconis, A Sly

342

2011

[A new coefficient of correlation](#)

S Chatterjee
Journal of the American Statistical Association, 1-14

298

2020

[Fluctuations of eigenvalues and second order Poincaré inequalities](#)

S Chatterjee
Probability Theory and Related Fields 143 (1), 1-40

247

2009

[The large deviation principle for the Erdős-Rényi random graph](#)

246

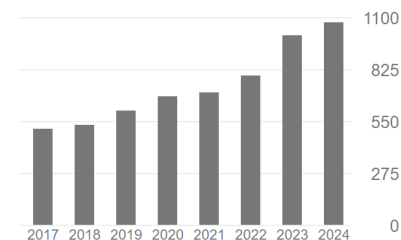
2011

引用先

[すべて表示](#)

すべて 2019 年以来

引用	7918	4893
h 指標	41	33
i10 指標	83	74



オープン アクセス

[すべて表示](#)

0 件の論文	38 件の論文
利用不可	利用可能

二変数 (X, Y) の間の非線形な関係を測る係数の提案

この係数は次のような特性をもつ

- Pearsonの相関係数や、Spearmanの順位相関係数などに代表される古典的な係数と同じくらいシンプル
- 変数間の依存度を測れる解釈可能な指標
- 独立性の仮定の下で、簡単な漸近理論が存在

Background:

二変数(X, Y) の関係性の強さをどうやって測るか??

古典的な係数

- Pearson's correlation
- Spearman's rank correlation, ρ
- Kendall's rank correlation, τ

⋮

- ✓ シンプル
- ✓ 依存度合いを解釈可能な形で測れる
- ✓ p値を計算するための漸近理論がある
- ⚠ 線形 or 単調な関係しか測れない

独立性検定

- Maximal information coefficient
- Distance correlation
- Heller-Heller-Gorfine (HHG) test
- Hilbert-Schmidt independence criterion

⋮

- ✓ 非線形な関係を検出可能
- ⚠ (X, Y) の関係の”強さ”は測れない
- ⚠ 計算コストが高い
漸近理論がないため、permutation test/bootstrapが必要

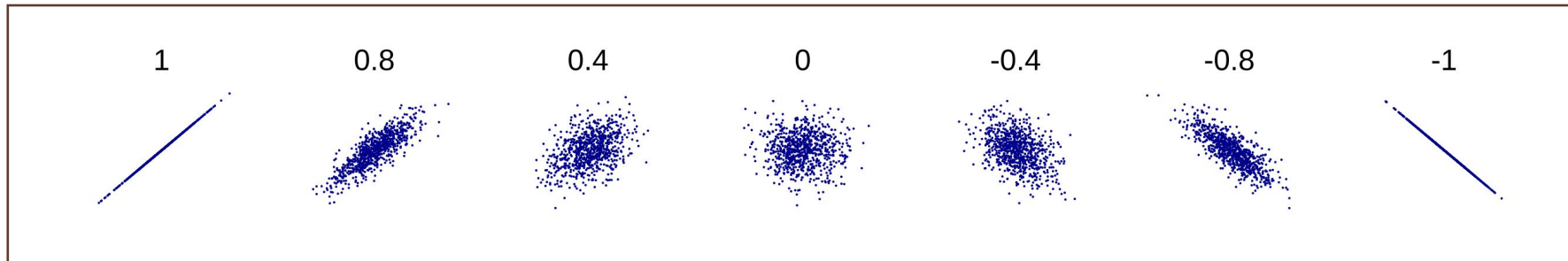
Background: Pearson correlation coefficient

(X, Y) の線形な関係の強さを測る

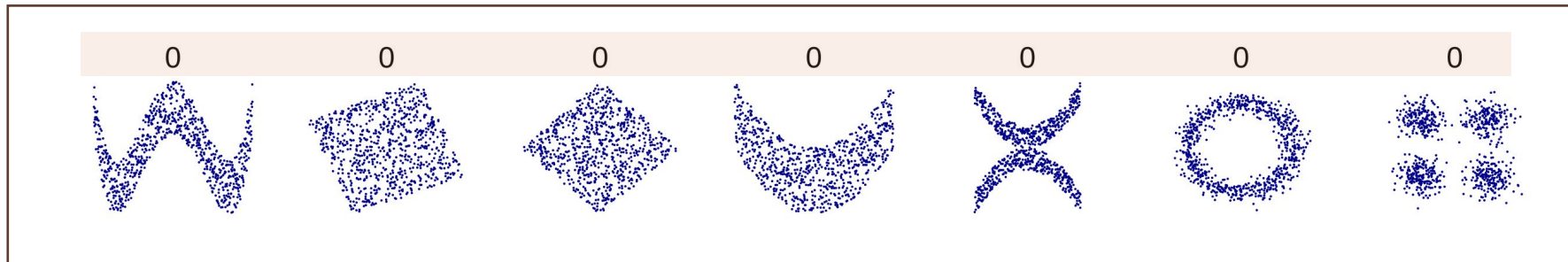
$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\sigma_X^2} \sqrt{\sigma_Y^2}} \cdot \quad \begin{array}{l} \text{COV : covariance} \\ \sigma_x^2, \sigma_y^2 : \text{variance} \end{array}$$

値の範囲: $-1 \leq \rho \leq 1$

✓ Linear



! Non-linear



Background: Spearman's rank

(X, Y) の単調な関係の強さを測る

(X_i, Y_i) をランキング $R[X_i], R[Y_i]$ に変換

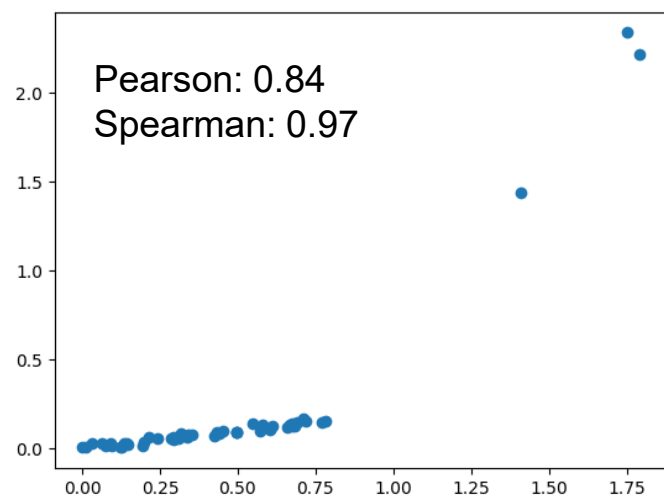
$$\rho = \frac{\text{cov}(R[X], R[Y])}{\sqrt{\sigma_{R[X]}^2} \sqrt{\sigma_{R[Y]}^2}}.$$
$$-1 \leq \rho \leq 1$$

If no tie value

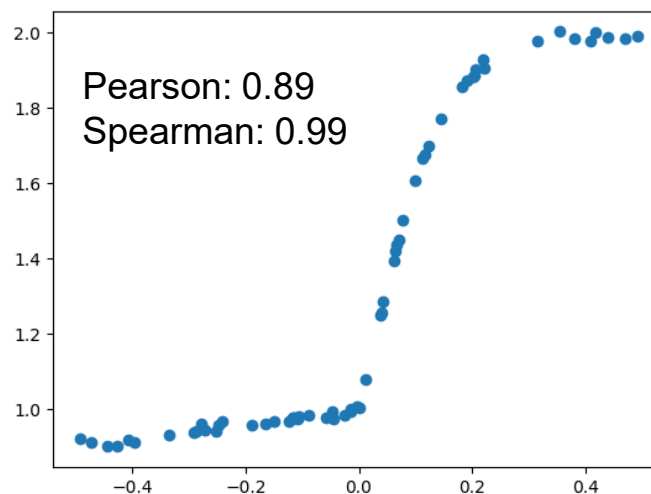
$$\rho = 1 - \frac{6 \sum_i^n d_i^2}{n(n^2 - 1)}$$
$$d_i = R[Y_i] - R[X_i]$$

X	Y	R[X]	R[Y]	d
0.13	0.29	5	5	0
0.75	0.41	1	4	-3
0.70	0.92	2	1	1
0.14	0.47	4	3	1
0.33	0.63	3	2	-1

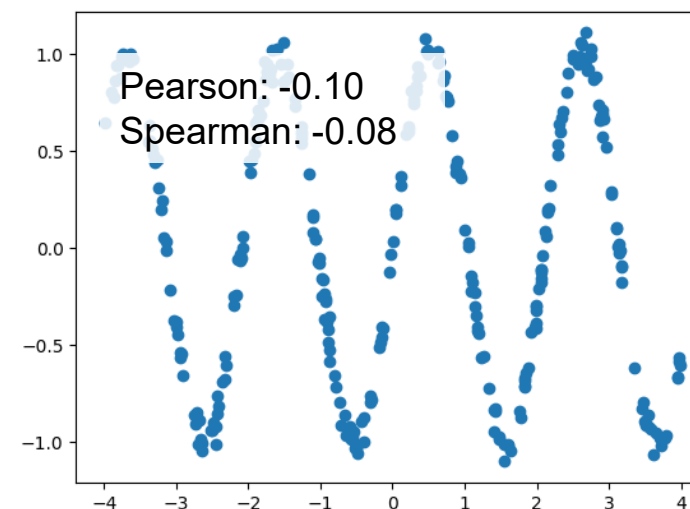
✓ Outlier



✓ Monotonic



⚠ Non-Linear



Background: Hilbert-Schmidt Independence Criterion (HSIC)

Roughly Speaking: データを再生核ヒルベルト空間に射影し、共分散を計算

✓ 独立性検定(パラメトリック) をもちいて非線形な関係を検出

1. 仮説の設定:

- 帰無仮説 H_0 : (X, Y) は独立
- 対立仮説 H_1 : (X, Y) は独立じゃない

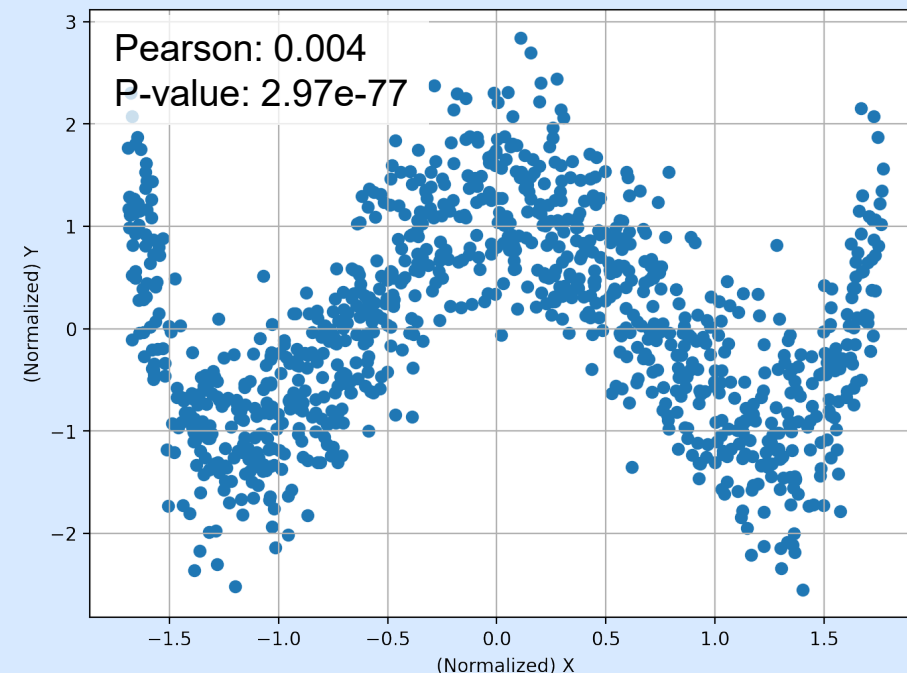
2. 優位水準 α (e.g. 0.05=95th percentile) を設定

HSIC の統計量は独立性の仮定の下で、ガンマ分布に従う(漸近理論)

3. HSIC統計量からp値を計算

4. 結論の導出

- もし、 p 値 < 優位水準:
 - 帰無仮説 H_0 を棄却し、 (X, Y) は独立ではない



⚠ 統計量を見ただけでは解釈が困難

⚠ 計算コストが高い

- HSICは(かなり荒い)漸近理論がたまたまあるが、ほとんどすべての手法は漸近理論がない
- このためPermutation testsやbootstrapといったノンパラメトリック検定が必要

この研究では、二変数 (X, Y) の**非線形な関係を測る新しい係数**を提案

- 古典的な係数と同じぐらい**シンプル**
- 変数間の依存度を測れる**解釈可能性が高い**指標
 - 値が $[0, 1]$ に制限される
 - 0であれば、 (X, Y) は独立
 - 1であれば可測関数 $Y = f(X)$
- 帰無仮説のもとで単純な**漸近理論**が存在
 - パラメトリック検定で独立性の検出が可能

Proposed Method: New coefficient ξ_n

0. n 個の*i.i.d* でサンプルされたペアを用意

$(X, Y): (X_1, Y_1), \dots, (X_n, Y_n), Y$ is not const.

1. X を基準に昇順にソート $X_{(1)} \leq \dots \leq X_{(n)}$

➡ $(X_{(1)}, Y_{(1)}), \dots, (X_{(n)}, Y_{(n)})$

2. Y を昇順にランキング付け

➡ $r_1, \dots, r_n$

例:

$(X, Y): (\underline{1.2}, 1.9), (\underline{0.3}, 1.1), (\underline{3.2}, 1.7)$

↓ Sort

$(X_{(i)}, Y_{(i)}): (0.3, \underline{1.1}), (1.2, \underline{1.9}), (3.2, \underline{1.7})$

↓ Rank

$r: 1, 3, 2$

3. 係数 ξ_n を計算

X_i, Y_i に同値がないとき

$$\xi_n(X, Y) := 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}.$$

同値があるとき:

Y を昇順にランク付けした l_i を用意して、

$$\xi_n(X, Y) := 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^n l_i (n - l_i)}.$$

Proposed Method: Theorem 1.1

$$\xi_n(X, Y) := 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}.$$

ここで、 $n \rightarrow \infty$ のとき、 $\xi_n(X, Y)$ は次式に収束する

$$\xi(X, Y) := \frac{\int \text{Var}(\mathbb{E}(1_{\{Y \geq t\}} | X)) d\mu(t)}{\int \text{Var}(1_{\{Y \geq t\}}) d\mu(t)}, \quad \mu \text{ is the law of } Y.$$

 $\xi(X, Y)$ は $[0, 1]$

$\xi(X, Y) \in [0, 1]$ For every t , $\text{Var}(1_{\{Y \geq t\}}) \geq \text{Var}(\mathbb{E}(1_{\{Y \geq t\}} | X))$

$\xi(X, Y) = 0$ X と Y が独立のとき, $\mathbb{E}(1_{\{Y \geq t\}} | X) = \text{constant}$

$\xi(X, Y) = 1$ Y が X の可測関数の時, $\text{Var}(\mathbb{E}(1_{\{Y \geq t\}} | X)) = 1_{\{Y \geq t\}}$

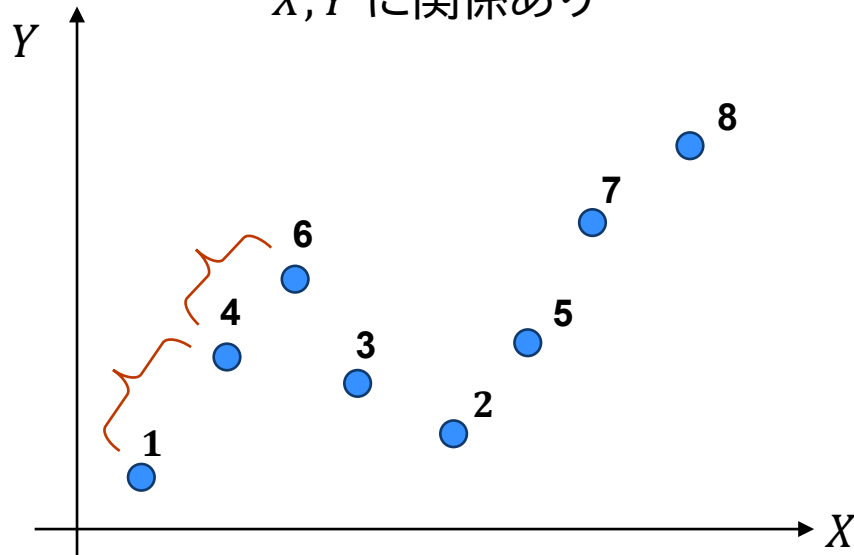
直感的な理解

X がわかったときに, Y をどれだけ説明できるか?

$$\xi_n(X, Y) := 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}.$$

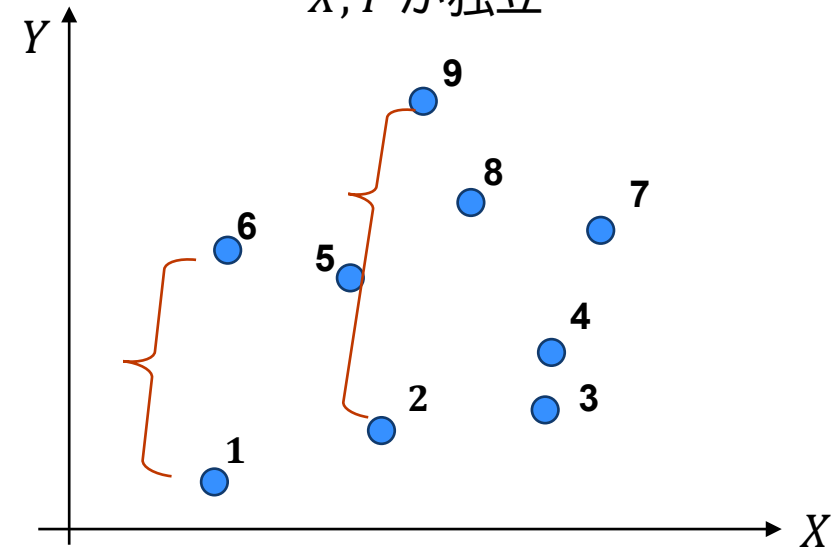
$(X, Y) := (X_1, Y_1), \dots, (X_n, Y_n)$ が与えられたとき、 X を基準にデータを整列

X, Y に関係あり



隣接するランク差
 $|r_{i+1} - r_i|$ が小さい $\Rightarrow \xi_n$ は1に近づく

X, Y が独立

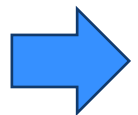


$|r_{i+1} - r_i|$ が大きい $\Rightarrow \xi_n$ は0に近づく

Proposed Method: Testing independence

独立性の仮定のもとで簡単な漸近理論が存在

Theorem 2.1. *Suppose that X and Y are independent and Y is continuous. Then $\sqrt{n}\xi_n(X, Y) \rightarrow N(0, 2/5)$ in distribution as $n \rightarrow \infty$.*

 $\xi_n(X, Y)$ を独立性検定につかえる

同様に Y が連続でないとき、 $n \rightarrow \infty$ で $\sqrt{n}\xi_n(X, Y) \rightarrow N(0, \tau^2)$ に従う

$$\tau^2 = \frac{\mathbb{E}\phi(Y_1, Y_2)^2 - 2\mathbb{E}(\phi(Y_1, Y_2)\phi(Y_1, Y_3)) + (\mathbb{E}\phi(Y_1, Y_2))^2}{(\mathbb{E}G(Y)(1 - G(Y)))^2},$$

$$F(t) := \mathbb{P}(Y \leq t), G(t) := \mathbb{P}(Y \geq t), \phi(y, y') := \min\{F(y), F(y')\}$$

Experiment 1: General Performance

シミュレーションデータの条件

- 3 種類の $Y = f(X)$
 - Linear
 - Parabolic
 - Sinusoid
- Noise を3つの強度で加える
- サンプルサイズ, $n = 100$

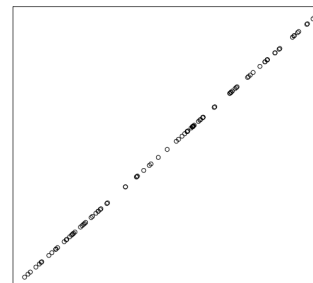
結果

- 大体関係を捉えられている.
 - $n = 100$ のとき、95thパーセンタイルは $\xi_n(X, Y) = 0.066$
 - すべての結果の ξ_n はそれよりもずっと大きい
- ξ_n は公平な(*equitable*) 係数に見える
 - 同じ強度のノイズを与えるとどれも同じようなスコアになっている
 - 数学的に厳密ではないが、直感的

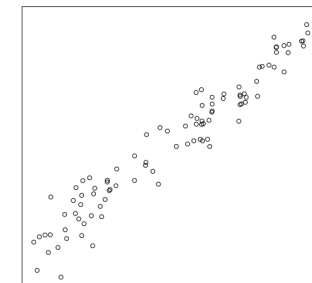
Noise increase



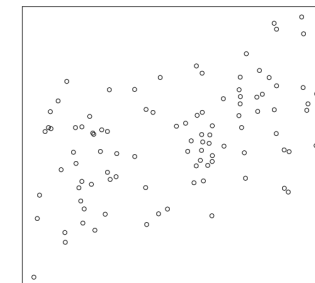
Linear



(a) $\xi_n = 0.970$.



(b) $\xi_n = 0.732$.

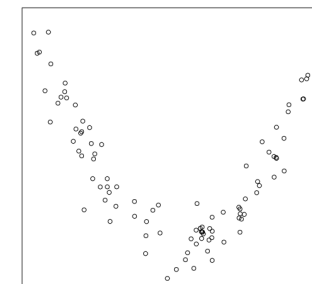


(c) $\xi_n = 0.145$.

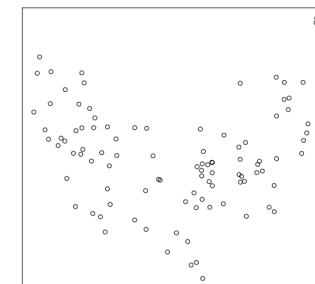
Para



(d) $\xi_n = 0.941$.

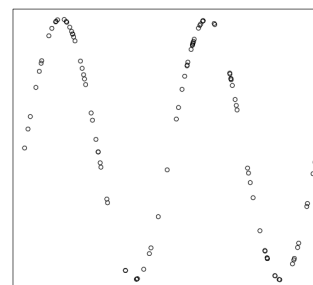


(e) $\xi_n = 0.684$.

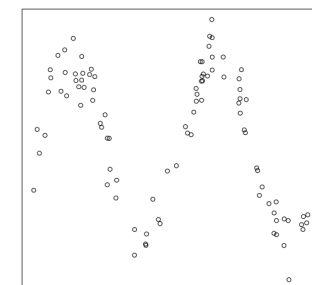


(f) $\xi_n = 0.265$.

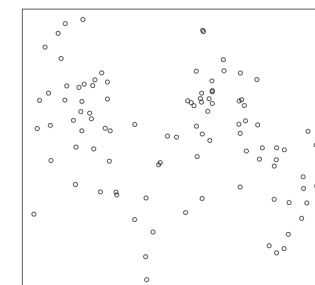
Sin



(g) $\xi_n = 0.885$.



(h) $\xi_n = 0.650$.



(i) $\xi_n = 0.281$.

Experiment 2:

Testing independence (Statistical power)

検出力: 帰無仮説 H_0 が正しくないとき、 H_0 を正しく棄却する確率

比較手法(一般的な検定量)

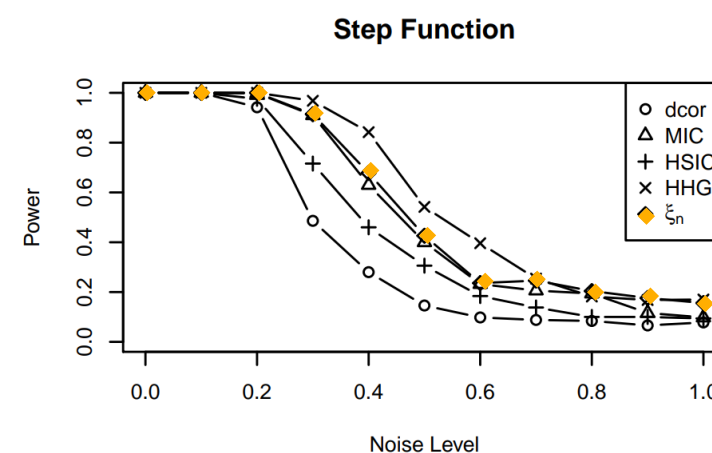
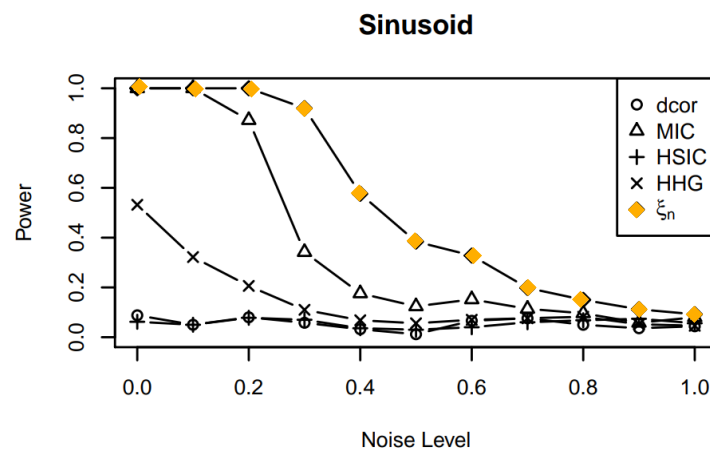
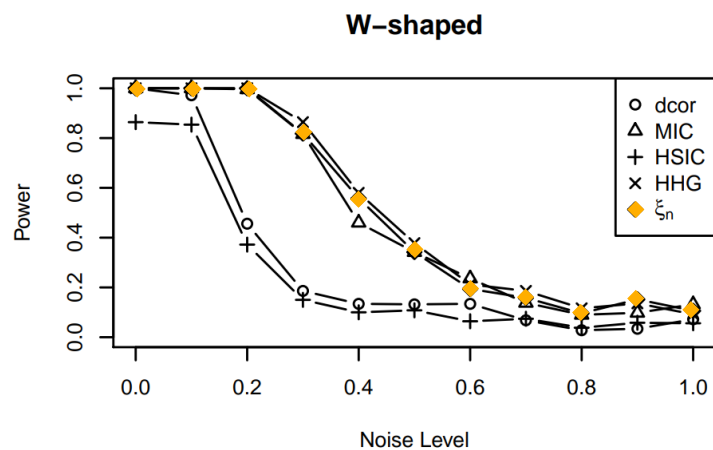
- Maximal information coefficient (MIC)
- Distance correlation (dcor)
- HHG test
- Hilbert-Schmidt independence criterion (HSIC)

シミュレーションデータ ($n = 100$, noise $\varepsilon \sim \mathcal{N}(0, 1)$, noise parameter $\lambda \in [0, 1]$)

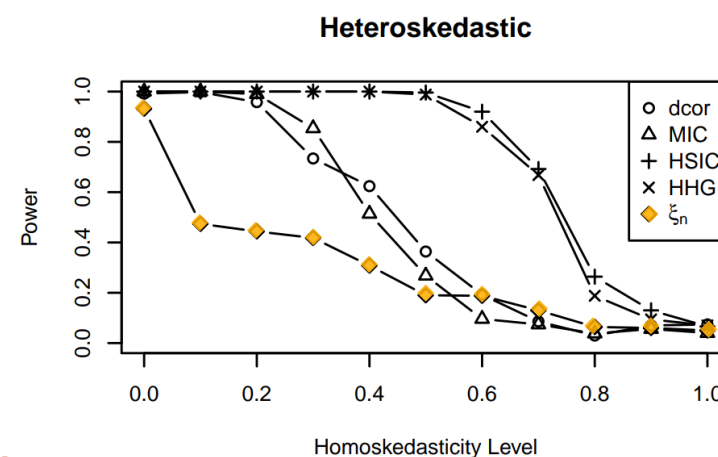
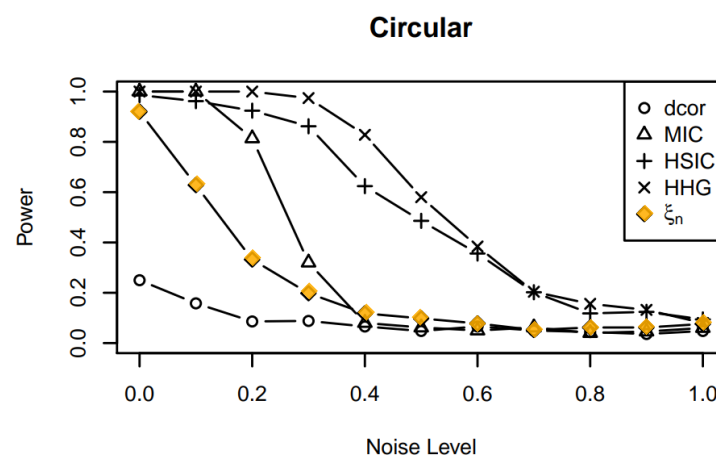
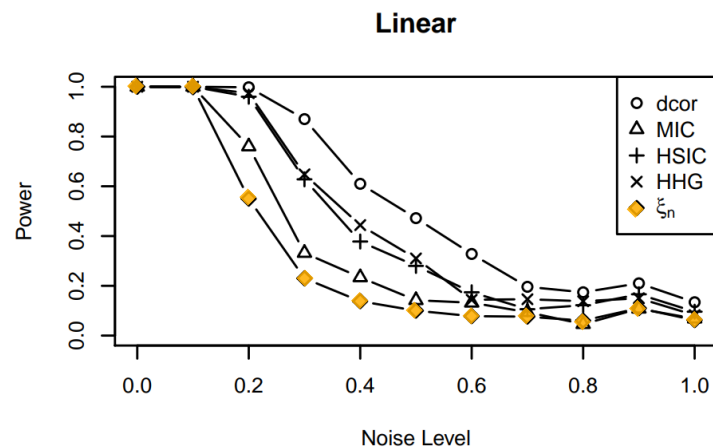
- Linear: $Y = 0.5X + 3\lambda\varepsilon$
- Step function: $Y = f(X) + 10\lambda\varepsilon$
- W-shaped, $Y = |X + 0.5|1_{\{X < 0\}} + |X - 0.5|1_{\{X \geq 0\}} + 0.75\lambda\varepsilon$
- Sinusoid: $Y = \cos 8\pi X + 3\lambda\varepsilon$
- Circular: $Y = Z\sqrt{1 - X^2} + 0.9\lambda\varepsilon$, where Z is 1 or -1 (random)
- Heteroskedastic: $Y = 3(\sigma(X)(1 - \lambda) + \lambda)\varepsilon$, where $\sigma(X) = 1$ if $|X| \leq 0.5$ and 0 otherwise
(不均一分散)

Experiment 2: 検出力

いい性能: ノイズレベルが上がったときに、検出力(Power)が落ちない



振動するような信号に対して ξ_n は検出力が強い



滑らかなデータに対しては、パフォーマンスが劣る

Experiment 3: Computational cost

Comparison of run times (in seconds)

n	dCor	MIC	HSIC	HHG	ξ_n
100	0.008	0.328	0.048	0.167	0.006
500	0.104	5.433	1.214	4.671	0.007
1000	0.532	17.459	5.028	20.515	0.009
2000	2.423	55.556	18.873	108.949	0.009
10000	88.976	1097.483	860.605	> 30 mins	0.011



Permutation test (200 permutations)

漸近理論

提案する新しい係数は様々な利点が存在する

1. とてもシンプルな式
 - 理解が簡単,
 - 計算が速い
2. ランクベースの手法のため、外れ値に頑健で、単調変換に対して不変
3. 値の範囲が $[0, 1]$.
 - $Y = f(X) \rightarrow \xi = 1$
 - X, Y が独立 $\rightarrow \xi = 0$
4. Y が定数ではないという以外の、 (X, Y) に対する仮定を必要としない
 - ξ_n はカテゴリカルデータにも適用できる
5. ξ_n は特に振動するような信号に対する検出力が高い

唯一の欠点: ほかの一般的な独立性検定手法に比べて、なめらかであったり振動しないような信号の場合には検出力が劣る

- 数式が煩雑で証明の理解はおいつかない
 - ξ の計算式が唐突にでてきたように見えるが、どうもよく調べると、[H.Dette et.al., 2013]によって提案されたコピュラベースの手法を、ランクベースで楽に解けるようにしたって話みたい
- 条件付き($T(Y, Z|X)$)の場合もChatterjeeによってのちに提案されている
- 利用できる場所はいろいろありそう。後続研究ではこのあたりの分野でよく使われていそう
 - 因果
 - 特徴量選択

[\[H.Dette et.al. 2013\] A Copula-Based Non-parametric Measure of Regression Dependence](#)

Appendix: Subsequent researches

- Azadkia and Chatterjee (2021). “A simple measure of conditional dependence”
 - Conditional Dependence Coefficient measures if Y is a function of $Z|X$ (represented at $T(Y, Z|X)$)
 - Use for Feature Selection

- Lin and Han(2021) “On boosting the power of Chatterjee's rank correlation”
 - Modified Chatterjee’s rank correlation.
 - Overcome the “Only one Disadvantage”

Only one Disadvantage: less power than several popular tests of independence when the signal is smooth and non-oscillatory

Appendix: Remarks

Maximum and minimum possible value of $\xi_n(X, Y)$

If there are no ties among the Y_i 's

$$\text{Max: } \frac{n-2}{n+1}$$

This can be less than 1 for small n .

e.g. $n = 20$, maximum possible value is 0.86

$$\text{Min: } -\frac{1}{2} + O(1/n)$$

This can be less than 0 !

Theorem 1.1 only applies to i.i.d samples

When a negative value of ξ_n : the data does not resemble an i.i.d sample.