A new coefficient of correlation

Sourav Chatterjee
The three most popular classical measures of statistical association are Pearson’s correlation coefficient, Spearman’s $\rho$, and Kendall’s $\tau$.

These coefficients are powerful for detecting linear or monotone associations, and they have well-developed asymptotic theories for calculating P-values.

However, a serious problem is that they are not effective for detecting associations that are not monotonic, even in the complete absence of noise.

There have been many proposals to address this deficiency of the classical coefficients, such as

- the maximal correlation coefficient,
- various coefficients based on joint cumulative distribution functions and ranks,
- kernel-based methods,
- information theoretic coefficients,
- coefficients based on copulas, and
- coefficients based on pairwise distances.
Some of these coefficients are popular among practitioners. But there are two common problems.

First, most of these coefficients are designed for testing independence, and not for measuring the strength of the relationship between the variables.

Ideally, one would like a coefficient that approaches its maximum value if and only if one variable looks more and more like a noiseless function of the other.

It is sometimes believed that the maximal information coefficient and the maximal correlation coefficient measure the strength of the relationship in the above sense, but that’s not correct.

Although MIC and maximal correlation are maximized when one variable is a function of the other, the converse is not true. They may be equal to 1 even if the relationship is very noisy. (An example is given in the paper.)
The second problem is that none of the coefficients for testing independence have simple asymptotic theories under the hypothesis of independence that facilitate the quick computation of P-values for testing independence.

In the absence of such theories, the only recourse is to use computationally expensive permutation tests or other kinds of bootstrap.
Goal of this talk

- One may wonder if it is at all possible to define a coefficient that is
  - as simple as the classical coefficients, and yet
  - is a consistent estimator of some measure of dependence which is 0 if
    and only if the variables are independent and 1 if and only if one is a
    measurable function of the other, and
  - has a simple asymptotic theory under the hypothesis of independence,
    like the classical coefficients.

- I will now present such a coefficient.

- The formula is so simple that it is likely that there are many such
  coefficients, some of them possibly having better properties than the
  one I am going to present.

- Reference:
A new coefficient of correlation

- Let \((X, Y)\) be a pair of random variables, where \(Y\) is not a constant.
- Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be i.i.d. pairs with the same law as \((X, Y)\), where \(n \geq 2\).
- The new coefficient has a simpler formula if the \(X_i\)'s and the \(Y_i\)'s have no ties. Let me present this simpler formula first, and then the general case.
- Suppose that the \(X_i\)'s and the \(Y_i\)'s have no ties.
- Rearrange the pairs as \((X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})\) such that \(X_{(1)} \leq \cdots \leq X_{(n)}\).
- Let \(r_i\) be the rank of \(Y_{(i)}\), that is, the number of \(j\) such that \(Y_{(j)} \leq Y_{(i)}\). The new correlation coefficient is defined as:

\[
\xi_n(X, Y) := 1 - \frac{3}{n^2 - 1} \sum_{i=1}^{n-1} |r_{i+1} - r_i|.
\]
In the presence of ties, $\xi_n$ is defined as follows.

If there are ties among the $X_i$’s, then choose an increasing rearrangement as before by breaking ties uniformly at random.

Let $r_i$ be as before, and additionally define $\ell_i$ to be the number of $j$ such that $Y(j) \geq Y(i)$.

Then define:

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

When there are no ties among the $Y_i$’s, $\ell_1, \ldots, \ell_n$ is just a permutation of 1, $\ldots$, $n$, and so the denominator in the above expression is just $n(n^2 - 1)/3$, which reduces this definition to the earlier expression.
The following theorem shows that $\xi_n$ is a consistent estimator of a certain measure of dependence between the random variables $X$ and $Y$.

**Theorem (C., 2021)**

*If $Y$ is not almost surely a constant, then as $n \to \infty$, $\xi_n(X, Y)$ converges almost surely to the deterministic limit*

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

*where $\mu$ is the law of $Y$. This limit belongs to the interval $[0, 1]$. It is 0 if and only if $X$ and $Y$ are independent, and it is 1 if and only if there is a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$ almost surely.*
Remarks

- Unlike most coefficients, $\xi_n$ is not symmetric in $X$ and $Y$.
- But we would like to keep it that way because we may want to understand if $Y$ is a function $X$, and not just if one of the variables is a function of the other.
- A symmetric measure of dependence, if required, can be easily obtained by taking the maximum of $\xi_n(X, Y)$ and $\xi_n(Y, X)$.
- By the theorem, this symmetrized coefficient converges in probability to $\max\{\xi(X, Y), \xi(Y, X)\}$, which is 0 if and only if $X$ and $Y$ are independent, and 1 if and only if at least one of $X$ and $Y$ is a measurable function of the other.
In the theorem, there are no restrictions on the law of \((X, Y)\) other than that \(Y\) is not a constant.

In particular, \(X\) and \(Y\) can be discrete, continuous, light-tailed or heavy-tailed.

The coefficient \(\xi_n(X, Y)\) remains unchanged if we apply strictly increasing transformations to \(X\) and \(Y\), because it is based on ranks.

For the same reason, it can be computed in time \(O(n \log n)\). (The actual computation on a computer is also very fast.)
The limiting value $\xi(X, Y)$ has appeared earlier in the literature (Dette et al. 2013, and Gamboa et al. 2018).

Dette et al. gave a copula-based estimator for $\xi(X, Y)$ when $X$ and $Y$ are continuous, that is consistent under smoothness assumptions on the copula and appears to be computable in time $n^{5/3}$ for an optimal choice of tuning parameters.

The coefficient $\xi_n$ looks similar to some coefficients defined earlier (e.g., by Friedman and Rafsky 1983), but in spite of its simple form, it seems to be genuinely new.
Asymptotic distribution under independence

- The main purpose of $\xi_n$ is to provide a measure of the strength of the relationship between $X$ and $Y$, and not to serve as a test statistic for testing independence.
- However, one can use it for testing independence if so desired. It has a nice and simple asymptotic theory under independence.
- The following theorem gives the asymptotic distribution of $\sqrt{n} \xi_n$ under the hypothesis of independence and the assumption that $Y$ is continuous. The more general asymptotic theory in the absence of continuity will be presented in the subsequent slides.

**Theorem (C., 2021)**

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 $n \rightarrow \infty$.

- The above result follows from an old result of Chao et al. (1993).
- In numerical examples, it is seen that the CLT is roughly valid even for $n$ as small as 20.
Asymptotic distribution under independence, contd.

- If $X$ and $Y$ are independent but $Y$ is not continuous, then also $\sqrt{n}\xi_n$ converges in distribution to a centered normal distribution, but the variance has a more complicated expression, and may depend on the law of $Y$.
- For each $t \in \mathbb{R}$, let $F(t) := \mathbb{P}(Y \leq t)$ and $G(t) := \mathbb{P}(Y \geq t)$. Let $\phi(y, y') := \min\{F(y), F(y')\}$, and define

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

where $Y_1, Y_2, Y_3$ are independent copies of $Y$. Then, we have:

**Theorem (C., 2021)**

Suppose that $X$ and $Y$ are independent. Then $\sqrt{n}\xi_n(X, Y)$ converges to $N(0, \tau^2)$ in distribution as $n \to \infty$, where $\tau^2$ is given by the formula stated above. The number $\tau^2$ is strictly positive if $Y$ is not a constant, and equals $2/5$ if $Y$ is continuous.
How to estimate $\tau^2$

- There is a simple way to estimate $\tau^2$ from the data using the estimator

$$\hat{\tau}^2_n = \frac{a_n - 2b_n + c_n^2}{d_n^2},$$

where $a_n$, $b_n$, $c_n$ and $d_n$ are defined as follows.

- For each $i$, let

$$R(i) := \#\{j : Y_j \leq Y_i\}, \quad L(i) := \#\{j : Y_j \geq Y_i\}.$$  

- Let $u_1 \leq u_2 \leq \cdots \leq u_n$ be an increasing rearrangement of $R(1), \ldots, R(n)$. Let $v_i := \sum_{j=1}^{i} u_j$ for $i = 1, \ldots, n$. Define

$$a_n := \frac{1}{n^4} \sum_{i=1}^{n} (2n - 2i + 1)u_i^2, \quad b_n := \frac{1}{n^5} \sum_{i=1}^{n} (v_i + (n - i)u_i)^2,$$

$$c_n := \frac{1}{n^3} \sum_{i=1}^{n} (2n - 2i + 1)u_i, \quad d_n := \frac{1}{n^3} \sum_{i=1}^{n} L(i)(n - L(i)).$$

Then we have the following result.
Consistency of $\hat{\tau}^2_n$

Theorem (C., 2021)

*The estimator $\hat{\tau}^2_n$ can be computed in time $O(n \log n)$, and converges to $\tau^2$ almost surely as $n \to \infty$. „*
Some simulated examples

![Scatter plots with values of $\xi_n$](image)

**Figure:** Values of $\xi_n$ for various kinds of scatterplots, with $n = 100$. 
A real data example

- In a landmark paper, Spellman et al. (1998) studied the expressions of 6223 yeast genes with the goal of identifying genes whose transcript levels oscillate during the cell cycle.
- In lay terms, this means that the expressions were studied over a number of successive time points (23, to be precise), and the goal was to identify the genes for which the transcript levels follow an oscillatory pattern.
- This example illustrates the utility of correlation coefficients in detecting patterns, because the number of genes is so large that identifying patterns by visual inspection is out of the question.
P-values were obtained for each gene, and a set of significant genes were selected using the Benjamini–Hochberg FDR procedure, with the expected proportion of false discoveries set at 0.05.

It turned out that there were 215 genes that were selected by $\xi_n$ but by none of the other tests that have been used previously.

The figure in the next slide shows the transcript levels of the top 6 of these genes (that is, those with the smallest P-values). As the figure shows, these genes exhibit almost perfect oscillatory behavior — and yet, they were not selected by other tests.
Transcript levels of the top 6 genes selected by $\xi_n$
Recent developments

A lot of follow-up work has accumulated since the paper was posted on arXiv. The following are some of the main developments:

- **Power for testing independence.** The coefficient was found to have one shortcoming, which is that it has lower power than some other tests for testing independence. There have been a number of suggested improvements to mend this deficit. The main papers on this, so far, are the following.
**Multivariate extensions.** There have been several attempts at generalizing the coefficient to the multivariate setting. The main papers on this, so far, are the following.

**Extension to conditional dependence.** Based on the same general principles as $\xi_n$, a coefficient for measuring conditional dependence was proposed in the following paper.


The conditional dependence coefficient was used to propose a variable selection algorithm called FOCI, that has various advantages over existing procedures. For power analysis and some further developments, see:


**Application to causal inference.** FOCI was used for designing a new algorithm for causal inference in the following paper.

Proof sketch for $\xi_n \to \xi$

- For simplicity, assume that $X$ and $Y$ are continuous variables.
- Recall that $r_i$ is the rank of $Y(i)$, where $(X(1), Y(1)), \ldots, (X(n), Y(n))$ is a rearrangement of the data in increasing order of $X_i$'s.
- Recall that $\xi_n = 1 - \frac{3}{n^2 - 1} \sum_{i=1}^{n} |r_{i+1} - r_i|$.
- Note that $r_i/n \approx F(Y(i))$, where $F$ is the c.d.f. of $Y$.
- Thus, $\xi_n \approx 1 - \frac{3}{n} \sum_{i=1}^{n} |F(Y_i) - F(Y_{N(i)})|$, where $N(i)$ is the index $j$ such that $X_j$ is immediately to the right of $X_i$.
- $|F(x) - F(y)| = \int (1_{\{t \leq x\}} - 1_{\{t \leq y\}})^2 d\mu(t)$, where $\mu$ is the law of $Y$.
- Since $X_i \approx X_{N(i)}$, the random variables $Y_i$ and $Y_{N(i)}$ are approximately i.i.d. conditional on $X = (X_1, \ldots, X_n)$, which gives:
  \[ \mathbb{E}[(1_{\{t \leq Y_i\}} - 1_{\{t \leq Y_{N(i)}\}})^2 |X] \approx 2\operatorname{Var}(1_{\{t \leq Y_i\}} |X) = 2\operatorname{Var}(1_{\{t \leq Y_i\}} |X_i). \]
  This gives $\mathbb{E}(1_{\{t \leq Y_i\}} - 1_{\{t \leq Y_{N(i)}\}})^2 \approx 2\mathbb{E}[(\operatorname{Var}(1_{\{t \leq Y\}} |X)).$
  So, we get $\mathbb{E}|F(Y_i) - F(Y_{N(i)})| \approx \int 2\mathbb{E}[(\operatorname{Var}(1_{\{t \leq Y\}} |X))]d\mu(t)$.
- From this, it is easy to show $\mathbb{E}(\xi_n) \to \xi$. Using concentration inequalities, we then get $\xi_n \to \xi$. 

Sourav Chatterjee
A new coefficient of correlation
Proof sketch for properties of $\xi$

- Recall that
  $$\xi(X, Y) = \frac{\int \text{Var}(\mathbb{E}(1\{Y \geq t\} | X))d\mu(t)}{\int \text{Var}(1\{Y \geq t\})d\mu(t)},$$

where $\mu$ is the law of $Y$.

- Since $\text{Var}(\mathbb{E}(1\{Y \geq t\} | X)) \leq \text{Var}(1\{Y \geq t\})$ for every $t$, we have $\xi \in [0, 1]$.

- It is not hard to see that $\text{Var}(\mathbb{E}(1\{Y \geq t\} | X)) = \text{Var}(1\{Y \geq t\})$ if and only if $1\{Y \geq t\}$ is a measurable function of $X$.

- This holds for all $t$ in the support of $Y$ if and only if $Y$ is a measurable function of $X$.

- Similarly, $\text{Var}(\mathbb{E}(1\{Y \geq t\} | X)) = 0$ if and only if $1\{Y \geq t\}$ is independent of $X$.

- Again, this holds for all $t$ in the support of $Y$ if and only if $Y$ and $X$ are independent.

- This proves that $\xi = 0$ iff $X$ and $Y$ are independent, and $\xi = 1$ iff $Y$ is a measurable function of $X$. 

Sourav Chatterjee

A new coefficient of correlation