



Adjust Pearson's r to Measure Arbitrary Monotone Dependence

Prof. Xinbo AI

Dec. 2024



Adjust Pearson's r to Measure Arbitrary Monotone Dependence





Methods

Results

Conclusion

Pearson's *r*

Motivation

- proposed in the late 19th century (Pearson, 1896)
- has been one of the main tools for scientists and engineers to study bivariate dependence during the 20th century
- still goes strong in the 21st century (Puccetti, 2022)



Methods

Results

Conclusion





Methods

Results



Widespread belief: *Pearson's r is only for linear dependence*

- Introduction to the theory of statistics, 1974: Both the covariance and the correlation coefficient of random variables *X* and *Y* are measures of a linear relationship of *X* and *Y*
- All of statistics: a concise course statistical inference, 2004: If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.
- Theoretical statistics: topics for a core course 2010: The covariance between two variables might be viewed as a measure of the linear association between the two variables
- Probability and statistics, 2012: The covariance and correlation are attempts to measure that dependence, but they only capture a particular type of dependence, namely linear dependence



Methods

Results

Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

The widespread belief is a myth

■ van den Heuvel and Zhan (2022): Pearson's *r* should not be ruled out *a priori* for measuring nonlinear monotone dependence

Although this potential has been recognized, the specific approach to using Pearson's r for accurate measurement of nonlinear monotone dependence remains unresolved.

This is the issue that our paper aims to address!

THE AMERICAN EDITORICS. Taylor & Francis (4) more with the mill is 44-422 Shore Sectioned Same haloning/10. Table/Male11.1011.0021.2000442 B OPEN ADCESS R. iteration Myths About Linear and Monotonic Associations: Pearson's r, Spearman's p, and Kendall's r Edwin van den Hesvel** and Zhuozhao Zhan* "Inventive Nedicine and Epidemiology, Tutucal of Medicine, Boston Univentity, Ecotor, NA, "Department of Nathematics, and Computer Science, Earthouses University of Technology, Eindhinees, The Netherlands. ARTICLE HISTORY Peanon's correlation coefficient is considered a measure of linear association between bivariate random Perceived June 2008 Annapola & Norsky Start vertables X and Y. It is recommanded num to use it for other items of associations, indeed, for nonlinear monotonic associations alternative measures like Spearman's rank and Kendall's tau constation coefficients are considered more appropriate. These views or opinions on the estimation of association are strongly Revolution to make rooted in the statistical and other empirical sciences. After defining linear and monotonic associations, we execution, Monotonic execution, Pearson's will demonstrate that these opinions are incorrect. Planson's constation coefficient should not be ruled out a priori for measuring nonlinear monotonic associations. We will provide examples of practically relevant combilition coefficient; Tambes of Invariant distribution functions with nonlinear monotonic associations for which Pearwark Spearman's sank correlate prelation is preferred over Spearmant rank and Kendal's tau correlation in testing the dependency between X and Y. Alternatively, we will provide a famile of biveriate distributions with a linear association vees X and Y for which Spearman's same and Kendoli's tau are preferred over Pearson's correlation. Our enumerics show that existing users on linnar and monotonic associations are moths. **I.** Introduction These views or opinions on linear and monotonic associations are strongly rooted in statistics and many other (empir-Measuring dependency or an association between two comical) sciences. Intaitive argumentation is (mostly) based on timuous random variables X and Y has been discussed in the specific characteristics of these correlation coefficients. Pearson's literature for a long time. The three most commonly used mear connects directly to the canonical correlation parameter of stees are Prarooth r. Spearman's p. and Kendall's r correlation the bivariate normal distribution, for which the association coefficients (Marshall 1998; Chok 2010), which is not surprising, hetseen X and Y is repically linear. Pearson's 7 also represents since they have been developed more than a contary ago (Galton 1890, Pearson 1895; Spearman 1904; Fechner 1897; Keralali The standardized slope of the variable X in a linear regression analysis of X on Y and vice versa (Monroe and Stait 1933). 1938, Kruskall 1958). The current view is that Pearson's r is comidered a measure for favor association, while Spearman's a and Moreover, when Pearson's correlation coefficient is senal to Kendalls r an measures for measures for association (Spearman, one or ratios ore, either $Y = u + \beta X$ or $X = a + \beta T$ occurs, indicating a linear relation on the two random variables 1904 Monroe and Stait 1933; Elashoff and Dunhar 1972; Speed. 2011; Au et al. 2013; Path, Neubäuser, and Baston 2015; Aliman Spearman's a and Kendull's 7 are rank coerditions or measures

and Kesywinski 2015. Ltu 2019). In line with this interpretation, the use of Pearson's r has been strongly discouraged for forms. X-dimension coincides or co-occur with a positive change in of associations other than linear associations (Monroe and Statt - the Y-dimension. Furthermore, they are both transformation 1933; Kowalski 1972; Speed 2011). Furthermore, Spearman's p invariant, Spearman's p and Kendally 7 for X and Y are the and Kandally r have been opposed for linear or normal associations for $\varphi(X)$ and $\varphi(T)$, when φ and φ are both monotone tions due to the supersority of Pearson's r (Parlie 1980; Kowalski aucreasing or monotone decreasing functions (Kruskall 1996). 1972. Borkovel 2002; Generat and Verret 2005), although they may still be more appropriate in settings with outliers and die not automatically imply a prolopointion for its use in skewness (Gedeon and Hollkster 1987; Cheik 2010; Kamat 2012) amesoing dependency for continuum bivariate distributions De Winter et al. 2016).

of concordance. They measure how a nontive change in the Although these characterizations are perfectly correct, they Indeed, Rodgers and Nicewander (1998) provided 13 different

CONTRET I Main you the Internet On and Opportunity of Mathematics and Computer Transm. Redinners Deviced y of Technings, 48. The Netherlands O Supplementary materials for this article are available online. Please go to revected/taniles carefv765.

 ADD The resident: Public And Add with fearse to Twice of Twents force, 102 The is or Open Access a title distributed uniter the terms of the Twente Open Additional Additiona Additional Additional Additional Additional Additional Addition ReDerivation (contact http://continentiment.org/firment/liping-no-noil contenty-cont_and/a set altered, standarment, at hold span-in-any-set

Different bounds lead to different capture ranges

For covariance, we have three bounds:

Motivation

$$\left|\operatorname{cov}(X,Y)\right| \leq \sqrt{\operatorname{var}(X)\operatorname{var}(Y)} \leq \frac{1}{2}\left(\operatorname{var}(X) + \operatorname{var}(Y)\right) \leq \frac{1}{2}\left(\operatorname{var}(X) + \operatorname{var}(Y) + \left|\overline{X} - \overline{Y}\right|^2\right)$$

1st bound: (Pearson's *r*) $r(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \rightarrow \text{linear: } Y = \alpha X + \beta$

■ 2nd bound: (Additivity Coefficient) $r^+(X, Y) = \frac{\operatorname{cov}(X,Y)}{\frac{1}{2}(\operatorname{var}(X) + \operatorname{var}(Y))} \rightarrow \operatorname{Additive:} Y = \pm X + \beta$

Methods

Results

Conclusion

■ 3rd bound: (Concordance Coefficient) $r^{=}(X, Y) = \frac{\operatorname{cov}(X,Y)}{\frac{1}{2}(\operatorname{var}(X) + \operatorname{var}(Y) + |\bar{X} - \bar{Y}|^2)} \rightarrow$ Identical: $Y = \pm X$

By now, all the efforts have only led to looser bounds and measures with narrower capture ranges. Could we possibly explore breakthroughs by approaching the problem from *the opposite direction*, aiming to achieve **a tighter bound** and consequently, devise a new measure with **a broader capture range**? **YES!**

Methods

Results

Conclusion



New inequality tighter than Cauchy-Schwarz inequality

Cauchy–Schwarz inequality

For random variables X and Y, we have

 $\left| EXY \right| \le \sqrt{EX^2 EY^2}$

The equality holds if and only if $P(Y = \alpha X) = 1$

For samples *x* and *y* we have

 $\left|\left\langle x, y\right\rangle\right| \le \|x\| \|y\|$

The equality holds if and only if *x* and *y* are

linearly dependent, i.e., y = ax for some constant a.

Our inequality

For random variables *X* and *Y*, we have $|EXY| \le |EX^{\uparrow}Y^{\uparrow}| \le \sqrt{EX^{2}EY^{2}}$

The equality on the left holds if and only if *X* and *Y* are monotone dependent.

For samples x and y we have $|\langle x, y \rangle| \le |\langle x^{\uparrow}, y^{\uparrow} \rangle| \le ||x|| ||y||$

The equality on the left holds if and only if *x* and *y* are monotone dependent.

Methods

Conclusion



New inequality tighter than Cauchy-Schwarz inequality

Theorem 1. For random variables X and Y, we have

 $|EXY| \leq |EX^{\uparrow}Y^{\uparrow}| \leq \sqrt{EX^2EY^2}.$

The equality on the left holds if and only if X and Y are monotone dependent, and the equality on the right holds if and only if $Y \stackrel{d}{=} \alpha X$, with sgn $(EXY) = \text{sgn}(\alpha)$.

Here, $\stackrel{d}{=}$ *denotes equality in distribution, and* $EX^{\uparrow}Y^{\ddagger}$ *is defined as:*

 $EX^{\uparrow}Y^{\ddagger} = \begin{cases} EX^{\uparrow}Y^{\uparrow}, if \ EXY \ge 0\\ EX^{\uparrow}Y^{\downarrow}, if \ EXY < 0 \end{cases}$

Theorem 2. For samples x and y we have

 $|\langle x, y \rangle| \leq \left| \left\langle x^{\uparrow}, y^{\updownarrow} \right\rangle \right| \leq ||x|| ||y||.$

The equality on the left holds if and only if x and y are monotone dependent, and the equality on the right holds if and only if y is arbitrary permutation of ax, with $sgn(\langle x, y \rangle) = sgn(a)$.

Here, $\langle x^{\uparrow}, y^{\downarrow} \rangle$ is defined as:

$$\left\langle x^{\uparrow}, y^{\uparrow} \right
angle = \left\{ egin{array}{cc} \left\langle x^{\uparrow}, y^{\uparrow}
ight
angle, if & \left\langle x, y
ight
angle \geqslant 0 \ \left\langle x^{\uparrow}, y^{\downarrow}
ight
angle, if & \left\langle x, y
ight
angle < 0 \end{array}
ight.$$

Methods

Motivation

Results

Conclusion



New inequality tighter than Cauchy-Schwarz inequality

Corollary 1. For random variables X and Y, we have covariance inequality series as: $|\operatorname{cov}(X,Y)| \leq |\operatorname{cov}(X^{\uparrow},Y^{\uparrow})| \leq \sqrt{\operatorname{var}(X)\operatorname{var}(Y)}$

$$|\operatorname{cov}(X, Y)| \leq |\operatorname{cov}(X^{\vee}, Y^{\vee})| \leq \sqrt{\operatorname{var}(X)\operatorname{var}(Y)}$$

$$\leq \frac{1}{2}\left(\operatorname{var}(X) + \operatorname{var}(Y)\right)$$

$$\leq \frac{1}{2}\left(\operatorname{var}(X) + \operatorname{var}(Y) + \left|\bar{X} - \bar{Y}\right|^{2}\right)$$

The first equality holds if and only if X and Y are monotone dependent, and the second equality holds if and only if $Y \stackrel{d}{=} \alpha X + \beta$, with sgn (cov (X, Y)) = sgn (α) .

Corollary 2. For samples x and y, we have covariance inequality series as

$$\begin{aligned} |s_{x,y}| &\leq \left| s_{x^{\uparrow},y^{\ddagger}} \right| \leq \sqrt{s_x^2 s_y^2} \\ &\leq \frac{1}{2} \left(s_x^2 + s_y^2 \right) \\ &\leq \frac{1}{2} \left(s_x^2 + s_y^2 + \left| \bar{x} - \bar{y} \right|^2 \right) \end{aligned}$$

The first equality holds if and only if x and y are monotone dependent, and the second equality holds if and only if y is arbitrary permutation of ax + b, with $sgn(s_{x,y}) = sgn(a)$.

Methods

Conclusion



The proposed Rearrangement Correlation

■ The Rearrangement Correlation of random variables *X* and *Y* is defined as:

$$r^{\#}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{cov}(X^{\uparrow},Y^{\uparrow})}$$

■ The Rearrangement Correlation of samples *x* and *y* is defined as:

$$r^{\#}(x, y) = \frac{s_{x, y}}{\left|s_{x^{\uparrow}, y^{\downarrow}}\right|}$$

Proposition 1. For random variables X, Y, and samples x, y, the following hold:

- $|r^{\#}(X,Y)| \leq 1$ and the equality holds if and only if X and Y are monotone dependent.
- $|r^{\#}(x,y)| \leq 1$ and the equality holds if and only if x and y are monotone dependent.

Methods

Conclusion



The proposed Rearrangement Correlation

■ Covariance inequality series, correlation coefficients and their capture ranges



Results



A toy example

Methods

Motivation

Compared to Spearman's ρ , $r^{\#}$ has a higher resolution and is more accurate x = (4,3,2,1) $y_1 = (5,4,3,2.00) \rightarrow r^{\#}(x,y_1) = 1.00, \rho(x,y_1) = 1.00$ $y_2 = (5,4,3,3.25) \rightarrow r^{\#}(x,y_2) = 0.93, \rho(x,y_2) = 0.80$ $y_3 = (5,4,3,3.50) \rightarrow r^{\#}(x,y_3) = 0.85, \rho(x,y_3) = 0.80$ $y_4 = (5,4,3,3.75) \rightarrow r^{\#}(x,y_4) = 0.76, \rho(x,y_4) = 0.80$ $y_5 = (5,4,3,4.50) \rightarrow r^{\#}(x,y_5) = 0.38, \rho(x,y_5) = 0.40$

Obviously, y_1 and x behaves exactly in the same way, with their values getting small and small step by step. The behavior of y_2 , y_3 , y_4 , and y_5 are becoming more and more different from that of x. However, the ρ values are all the same for y_2 , y_3 and y_4 . In contrast, the $r^{\#}$ values can reveal all these differences exactly.

Motivation Methods

Results

Conclusion



Performance of different measures

■ in 50 simulated scenarios



Motivation Methods

Results

Conclusion



Performance of different measures

■ in 5 Real-life Scenarios





We may draw the conclusion that:

Pearson's *r* is undoubtedly the gold measure for linear dependence.

Now, it might be the gold measure also for nonlinear monotone dependence, if adjusted.





Thank you



