Robust Rank Correlation Coefficients on the Basis of Fuzzy Orderings

Vipin Kumar
MA20MSCST11002

Department of Mathematics
IIT Hyderabad

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Introduction

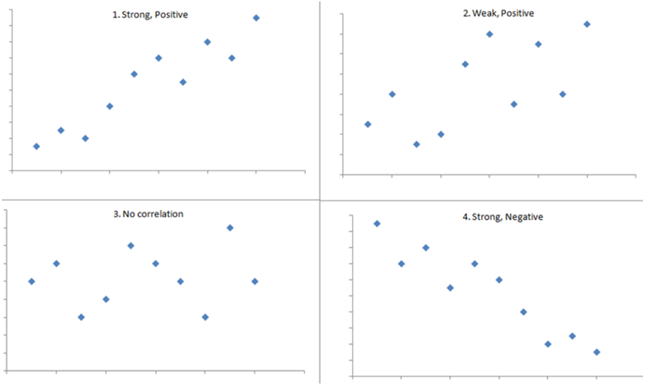
Correlation

Correlation is a statistical term describing the degree to which two variables move in coordination with one another.

Introduction

Correlation

Correlation is a statistical term describing the degree to which two variables move in coordination with one another.



Introduction

The most prominent representative is surely **Pearson's product moment coefficient** also called **correlation coefficient** for short. if $(x_i, y_i)_{i=1}^{n}$ is the data then correlation coefficient is

Introduction

The most prominent representative is surely **Pearson's product moment coefficient** also called **correlation coefficient** for short. if $(x_i, y_i)_{i=1}^{n}$ is the data then correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

Introduction

The most prominent representative is surely **Pearson's product moment coefficient** also called **correlation coefficient** for short. if $(x_i, y_i)_{i=1}^n$ is the data then correlation coefficient is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where

x_i, y_i are the values of x-variable, y-variable in a sample resp.

$\bar{x}, \bar{y}$ are the mean of the values of x-variable, y-variable resp.

Introduction

The most prominent representative is surely **Pearson's product moment coefficient** also called **correlation coefficient** for short. if $(x_i, y_i)_{i=1}^n$ is the data then correlation coefficient is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where

x_i, y_i are the values of x-variable, y-variable in a sample resp.

$\bar{x}, \bar{y}$ are the mean of the values of x-variable, y-variable resp.

Correlation coeff. is applicable to numerical data and assumes a linear relationship as the underlying model.

Introduction

The most prominent representative is surely **Pearson's product moment coefficient** also called **correlation coefficient** for short. if $(x_i, y_i)_{i=1}^n$ is the data then correlation coefficient is

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where

x_i, y_i are the values of x-variable, y-variable in a sample resp.
$\bar{x}, \bar{y}$ are the mean of the values of x-variable, y-variable resp.

Correlation coeff. is applicable to numerical data and assumes a linear relationship as the underlying model.

Drawback - It can not be used to detect non-linear relationship.

Introduction

Rank correlation is a method of finding the degree of association between two variables. The calculation for the rank correlation coefficient the same as that for the Pearson correlation coefficient,

Introduction

Rank correlation is a method of finding the degree of association between two variables. The calculation for the rank correlation coefficient the same as that for the Pearson correlation coefficient, but is calculated using the ranks of the observations and not their numerical values.

Introduction

Rank correlation is a method of finding the degree of association between two variables. The calculation for the rank correlation coefficient the same as that for the Pearson correlation coefficient, but is calculated using the ranks of the observations and not their numerical values. This method is useful when the data are not available in numerical form but information is sufficient to rank the data.

Introduction

Rank correlation is a method of finding the degree of association between two variables. The calculation for the rank correlation coefficient the same as that for the Pearson correlation coefficient, but is calculated using the ranks of the observations and not their numerical values. This method is useful when the data are not available in numerical form but information is sufficient to rank the data.

The two most common approaches are -

- Spearman's rank correlation coefficient (short Spearman's rho)

Introduction

Rank correlation is a method of finding the degree of association between two variables. The calculation for the rank correlation coefficient the same as that for the Pearson correlation coefficient, but is calculated using the ranks of the observations and not their numerical values. This method is useful when the data are not available in numerical form but information is sufficient to rank the data.

The two most common approaches are -

- Spearman's rank correlation coefficient (short Spearman's rho)
- Kendall's tau (rank correlation coefficient)

Spearman's rho

Spearman's rho

It measures the sum of quadratic distances of ranks and scales this measure to the interval [-1, 1].

Spearman's rho

Spearman's rho

It measures the sum of quadratic distances of ranks and scales this measure to the interval [-1, 1].

If we are given a family of pairs as (x_i, y_i) where all x_i and y_i are from linearly ordered domains X and Y resp. then Spearman's rho is computed as

Spearman's rho

Spearman's rho

It measures the sum of quadratic distances of ranks and scales this measure to the interval [-1, 1].

If we are given a family of pairs as (x_i, y_i) where all x_i and y_i are from linearly ordered domains X and Y resp. then Spearman's rho is computed as

$$\rho = 1 - 6\frac{\sum_{i=1}^{n}(r(x_i) - r(y_i))^2}{n(n^2 - 1)}$$

Spearman's rho

It measures the sum of quadratic distances of ranks and scales this measure to the interval [-1, 1].

If we are given a family of pairs as (x_i, y_i) where all x_i and y_i are from linearly ordered domains X and Y resp. then Spearman's rho is computed as

$$\rho = 1 - 6\frac{\sum_{i=1}^{n}(r(x_i) - r(y_i))^2}{n(n^2 - 1)}$$

where $r(x_i)$ is the rank of value x_i after sorting the list $(x_1, x_2, ..., x_n)$; $r(y_i)$ is defined analogously.

Spearman's rho

Spearman's rho

It measures the sum of quadratic distances of ranks and scales this measure to the interval [-1, 1].

If we are given a family of pairs as (x_i, y_i) where all x_i and y_i are from linearly ordered domains X and Y resp. then Spearman's rho is computed as

$$\rho = 1 - 6\frac{\sum_{i=1}^{n}(r(x_i) - r(y_i))^2}{n(n^2 - 1)}$$

where $r(x_i)$ is the rank of value x_i after sorting the list $(x_1, x_2, ..., x_n)$; $r(y_i)$ is defined analogously.

Note - It did not take coinciding values (ties) into account.

Kendall's tau

Before going to Kendall's tau we need to know about the concepts of concordance, discordance and ties.

Kendall's tau

Before going to Kendall's tau we need to know about the concepts of concordance, discordance and ties.

Concordant

For a given pair of indices $(i, j) \in \{1, 2, ..., n\}^2$,we say (i, j) is **concordant** if $x_i < x_j$ and $y_i < y_j$

Kendall's tau

Before going to Kendall's tau we need to know about the concepts of concordance, discordance and ties.

Concordant

For a given pair of indices $(i, j) \in \{1, 2, ..., n\}^2$,we say (i, j) is **concordant** if $x_i < x_j$ and $y_i < y_j$

$$C = |\{(i, j)|x_i < x_j \ and \ y_i < y_j\}|$$

Kendall's tau

Before going to Kendall's tau we need to know about the concepts of concordance, discordance and ties.

Concordant

For a given pair of indices $(i, j) \in \{1, 2, ..., n\}^2$,we say (i, j) is **concordant** if $x_i < x_j$ and $y_i < y_j$

$$C = |\{(i, j)|x_i < x_j \ and \ y_i < y_j\}|$$

Discordant

For a given pair of indices $(i, j) \in \{1, 2, ..., n\}^2$,we say (i, j) is **discordant** if $x_i < x_j$ and $y_i > y_j$

Kendall's tau

Before going to Kendall's tau we need to know about the concepts of concordance, discordance and ties.

Concordant

For a given pair of indices $(i, j) \in \{1, 2, ..., n\}^2$,we say (i, j) is **concordant** if $x_i < x_j$ and $y_i < y_j$

$$C = |\{(i, j)|x_i < x_j \ and \ y_i < y_j\}|$$

Discordant

For a given pair of indices $(i, j) \in \{1, 2, ..., n\}^2$,we say (i, j) is **discordant** if $x_i < x_j$ and $y_i > y_j$

$$D = |\{(i, j)|x_i < x_j \ and \ y_i > y_j\}|$$

Kendall's tau

Tie

If $x_i = x_j$, we say that (i, j) is a **tie in the first component**.

Kendall's tau

Tie

If $x_i = x_j$, we say that (i, j) is a **tie in the first component**. If $yi = yj$, we say that (i, j) is a **tie in the second component**.

Kendall's tau

Tie

If $x_i = x_j$, we say that (i, j) is a **tie in the first component**. If $yi = yj$, we say that (i, j) is a **tie in the second component**.We simply say that (i, j) is a **tie** if (i, j) is a tie in either component.

Kendall's tau

Tie

If $x_i = x_j$, we say that (i, j) is a **tie in the first component**. If $yi = yj$, we say that (i, j) is a **tie in the second component**.We simply say that (i, j) is a **tie** if (i, j) is a tie in either component.

$$T = |\{(i, j)|x_i = x_j\}|$$

Tie

If $x_i = x_j$, we say that (i, j) is a **tie in the first component**. If $yi = yj$, we say that (i, j) is a **tie in the second component**.We simply say that (i, j) is a **tie** if (i, j) is a tie in either component.

$$T = |\{(i, j)|x_i = x_j\}|$$

$$U = |\{(i, j)|y_i = y_j\}|$$

Kendall's tau

Tie

If $x_i = x_j$, we say that (i, j) is a **tie in the first component**. If $yi = yj$, we say that (i, j) is a **tie in the second component**.We simply say that (i, j) is a **tie** if (i, j) is a tie in either component.

$$T = |\{(i, j)|x_i = x_j\}|$$

$$U = |\{(i, j)|y_i = y_j\}|$$

The basic variant of Kendall's tau which we denote with τ_a is computed as -

$$\boxed{\tau_a = \frac{C - D}{\frac{1}{2}n(n-1)}}$$

Observation -

1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.

Kendall's tau

Observation -

1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.
2) If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value 1.

Kendall's tau

Observation -

1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.
2) If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value 1.
3) Ties are not counted in the τ_a definition.

Observation -

1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.
2) If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value 1.
3) Ties are not counted in the τ_a definition.
4) In extremal cases, Kendall's tau gives the same results as Spearman's rho.

Kendall's tau

Observation -

1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.
2) If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value 1.
3) Ties are not counted in the τ_a definition.
4) In extremal cases, Kendall's tau gives the same results as Spearman's rho.

> τ_a is best suited for detecting strictly monotonic relationships, but not ideally suited in the presence of ties.

A well-established second variant is the following:

Kendall's tau

A well-established second variant is the following:

$$\tau_b = \frac{C - D}{\sqrt{\frac{1}{2}n(n-1) - T}\sqrt{\frac{1}{2}n(n-1) - U}}$$

Kendall's tau

A well-established second variant is the following:

$$\tau_b = \frac{C - D}{\sqrt{\frac{1}{2}n(n-1) - T}\sqrt{\frac{1}{2}n(n-1) - U}}$$

It takes ties into account, but is still not fully robust to ties.

Kendall's tau

A well-established second variant is the following:

$$\tau_b = \frac{C - D}{\sqrt{\frac{1}{2}n(n-1) - T}\sqrt{\frac{1}{2}n(n-1) - U}}$$

It takes ties into account, but is still not fully robust to ties.

A simple and tie-robust rank correlation measure is the **gamma rank correlation** measure according to Goodman and Kruskal that is defined as

$$\gamma = \frac{C - D}{C + D}$$

Kendall's tau

A well-established second variant is the following:

$$\tau_b = \frac{C - D}{\sqrt{\frac{1}{2}n(n-1) - T}\sqrt{\frac{1}{2}n(n-1) - U}}$$

It takes ties into account, but is still not fully robust to ties.

A simple and tie-robust rank correlation measure is the **gamma rank correlation** measure according to Goodman and Kruskal that is defined as

$$\gamma = \frac{C - D}{C + D}$$

Remark - τ_a, τ_b and γ coincide in all cases where no ties occur in the data.

Motivation

Historically, all rank correlation measures highlighted above have been introduced with the aim to measure rank correlation of ordinal data (e.g. natural numbers, marks, quality classes, ranks).

Motivation

Historically, all rank correlation measures highlighted above have been introduced with the aim to measure rank correlation of ordinal data (e.g. natural numbers, marks, quality classes, ranks). But the measurement of rank correlation for real-valued data is equally important, but raises completely new issues.

Motivation

Historically, all rank correlation measures highlighted above have been introduced with the aim to measure rank correlation of ordinal data (e.g. natural numbers, marks, quality classes, ranks). But the measurement of rank correlation for real-valued data is equally important, but raises completely new issues.

The numerical data are almost always subject to **random perturbations (noise)** which the concepts introduced above do not take it into account.

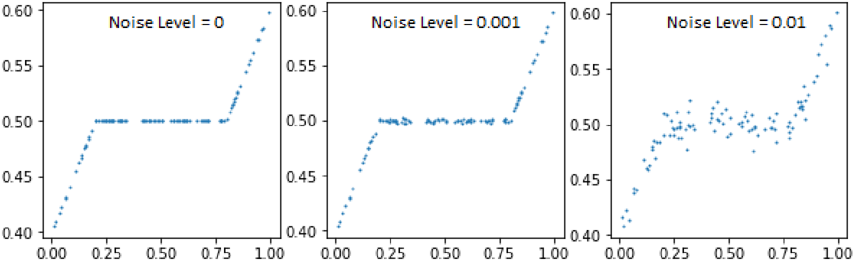
Motivation

Historically, all rank correlation measures highlighted above have been introduced with the aim to measure rank correlation of ordinal data (e.g. natural numbers, marks, quality classes, ranks). But the measurement of rank correlation for real-valued data is equally important, but raises completely new issues.

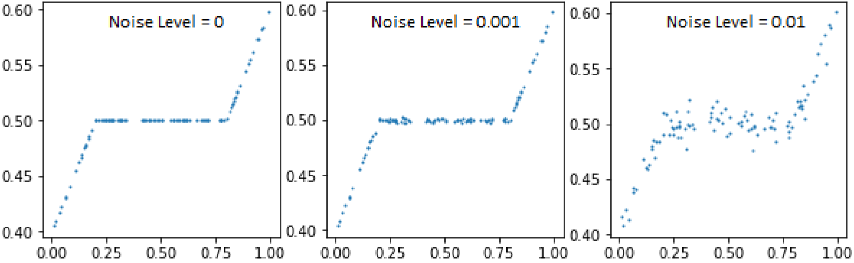
The numerical data are almost always subject to **random perturbations (noise)** which the concepts introduced above do not take it into account.

This is because pairs are concordant or discordant only on the basis of **ordering relations**, but without taking into account that only minimal differences may decide whether a pair is concordant or discordant.

Motivation


Noise Level = 0
Noise Level = 0.001
Noise Level = 0.01
0.60
0.55
0.50
0.45
0.40
0.00
0.25
0.50
0.75
1.00

Motivation



Noise level	0	0.001	0.01
ρ	0.8727	0.7730	0.7063
τ_a	0.61797	0.6254	0.5507
τ_b	0.7941	0.6318	0.5563
γ	1.0	0.6254	0.5507

So we can conclude that it is rather the **presence of noise** in general than the magnitude of noise that distracts the rank correlation measures.

Motivation

So we can conclude that it is rather the **presence of noise** in general than the magnitude of noise that distracts the rank correlation measures.

How we can define a **robust rank correlation** measure that depends continuously on the data by taking similarities into account, but still serves as a meaningful measure of rank correlation.

Fuzzy Orderings

Fuzzy Ordering

A fuzzy relation $L : X^2 \to [0, 1]$ is called **fuzzy ordering** with respect to a t-norm 'T' and a T-equivalence $E : X^2 \to [0, 1]$, for brevity T-E-ordering,

Fuzzy Orderings

Fuzzy Ordering

A fuzzy relation $L : X^2 \to [0, 1]$ is called **fuzzy ordering** with respect to a t-norm 'T' and a T-equivalence $E : X^2 \to [0, 1]$, for brevity T-E-ordering, if and only if the following three axioms for all $x, y, z \in X$:

Fuzzy Orderings

Fuzzy Ordering

A fuzzy relation $L : X^2 \to [0,1]$ is called **fuzzy ordering** with respect to a t-norm 'T' and a T-equivalence $E : X^2 \to [0,1]$, for brevity T-E-ordering, if and only if the following three axioms for all $x, y, z \in X$:

- E-Reflexivity: $E(x,y) \leq L(x,y)$

Fuzzy Ordering

A fuzzy relation $L : X^2 \to [0,1]$ is called **fuzzy ordering** with respect to a t-norm 'T' and a T-equivalence $E : X^2 \to [0,1]$, for brevity T-E-ordering, if and only if the following three axioms for all $x, y, z \in X$:

- E-Reflexivity: $E(x,y) \leq L(x,y)$
- T-E-antisymmetry: $T(L(x,y), L(y,x)) \leq E(x,y)$

Fuzzy Ordering

A fuzzy relation $L : X^2 \to [0,1]$ is called **fuzzy ordering** with respect to a t-norm 'T' and a T-equivalence $E : X^2 \to [0,1]$, for brevity T-E-ordering, if and only if the following three axioms for all $x, y, z \in X$:

- E-Reflexivity: $E(x,y) \leq L(x,y)$
- T-E-antisymmetry: $T(L(x,y), L(y,x)) \leq E(x,y)$
- T-transitivity: $T(L(x,y), L(y,z)) \leq L(x,z)$

Fuzzy Orderings

Fuzzy Ordering

A fuzzy relation $L : X^2 \to [0,1]$ is called **fuzzy ordering** with respect to a t-norm 'T' and a T-equivalence $E : X^2 \to [0,1]$, for brevity T-E-ordering, if and only if the following three axioms for all $x, y, z \in X$:

- E-Reflexivity: $E(x,y) \leq L(x,y)$
- T-E-antisymmetry: $T(L(x,y), L(y,x)) \leq E(x,y)$
- T-transitivity: $T(L(x,y), L(y,z)) \leq L(x,z)$

Moreover, we call a T-E-ordering **L strongly complete** if

$$max(L(x,y), L(y,x)) = 1$$

for all $x, y \in X$

Fuzzy Orderings

Assume $r > 0$,

Fuzzy Orderings

Assume $r > 0$, then

$$E_r(x, y) = max(0, 1 - \frac{1}{r}|x - y|)$$

is a T_L - equivalence on $\mathbb{R}$

Fuzzy Orderings

Assume $r > 0$, then

$$E_r(x, y) = max(0, 1 - \frac{1}{r}|x - y|)$$

is a T_L - equivalence on $\mathbb{R}$, where $T_L(x, y) = max(0, x + y - 1)$ denotes the lukasiewicz t-norm.

Fuzzy Orderings

Assume $r > 0$, then

$$E_r(x, y) = max(0, 1 - \frac{1}{r}|x - y|)$$

is a T_L - equivalence on $\mathbb{R}$, where $T_L(x, y) = max(0, x + y - 1)$ denotes the lukasiewicz t-norm.

For the above E_r and T_L we can prove that

$$L_r(x, y) = min(1, max(0, 1 - \frac{1}{r}(x - y)))$$

is a strongly complete $T_L - E_r$-ordering on $\mathbb{R}$.

Strict fuzzy ordering

A binary fuzzy relation $R : X^2 \to [0,1]$ is called **strict fuzzy ordering** with respect to T and a T-equivalence $E : X^2 \to [0,1]$, for brevity strict T-E-ordering,

Strict fuzzy ordering

Strict fuzzy ordering

A binary fuzzy relation $R : X^2 \to [0,1]$ is called **strict fuzzy ordering** with respect to T and a T-equivalence $E : X^2 \to [0,1]$, for brevity strict T-E-ordering, if it satisfy the following three properties

Strict fuzzy ordering

Strict fuzzy ordering

A binary fuzzy relation $R : X^2 \to [0, 1]$ is called **strict fuzzy ordering** with respect to T and a T-equivalence $E : X^2 \to [0, 1]$, for brevity strict T-E-ordering, if it satisfy the following three properties

- Irreflexive: $R(x, x) = 0$ for all $x \in X$

Strict fuzzy ordering

Strict fuzzy ordering

A binary fuzzy relation $R : X^2 \to [0,1]$ is called **strict fuzzy ordering** with respect to T and a T-equivalence $E : X^2 \to [0,1]$, for brevity strict T-E-ordering, if it satisfy the following three properties

- Irreflexive: $R(x,x) = 0$ for all $x \in X$
- T-transitive: $T(R(x,y), R(y,z)) \leq R(x,z)$ for all $x, y, z \in X$

Strict fuzzy ordering

A binary fuzzy relation $R : X^2 \to [0,1]$ is called **strict fuzzy ordering** with respect to T and a T-equivalence $E : X^2 \to [0,1]$, for brevity strict T-E-ordering, if it satisfy the following three properties

- Irreflexive: $R(x,x) = 0$ for all $x \in X$
- T-transitive: $T(R(x,y), R(y,z)) \leq R(x,z)$ for all $x,y,z \in X$
- E-extensional: $T(E(x,x^{'}), E(y,y^{'}), R(x,y)) \leq R(x^{'},y^{'})$ for all $x,x^{'},y,y',z \in X$

Strict fuzzy ordering

Strict fuzzy ordering

A binary fuzzy relation $R : X^2 \to [0,1]$ is called **strict fuzzy ordering** with respect to T and a T-equivalence $E : X^2 \to [0,1]$, for brevity strict T-E-ordering, if it satisfy the following three properties

- Irreflexive: $R(x,x) = 0$ for all $x \in X$
- T-transitive: $T(R(x,y), R(y,z)) \leq R(x,z)$ for all $x, y, z \in X$
- E-extensional: $T(E(x,x^{'}), E(y,y^{'}), R(x,y)) \leq R(x^{'},y^{'})$ for all $x, x^{'}, y, y', z \in X$

Given a T-E-ordering $L : X^2 \to [0,1]$,

$$R(x,y) = min(L(x,y), N_T(L(y,x))) \qquad (1)$$

Strict fuzzy ordering

where $N_T(x) = sup\{y \in [0,1] | T(x,y) = 0\}$ is the residual negation of T.

Strict fuzzy ordering

where $N_T(x) = sup\{y \in [0,1] | T(x,y) = 0\}$ is the residual negation of T. This is the most appropriate choice for extracting a strict fuzzy ordering from a given fuzzy ordering L.

Strict fuzzy ordering

where $N_T(x) = sup\{y \in [0,1] | T(x,y) = 0\}$ is the residual negation of T. This is the most appropriate choice for extracting a strict fuzzy ordering from a given fuzzy ordering L.

From this construction we can infer that the fuzzy relation

$$R_r(x,y) = min(1, max(0, \frac{1}{r}(y-x)))$$

is a T_L-E_r-ordering.

Strict fuzzy ordering

where $N_T(x) = sup\{y \in [0,1] | T(x,y) = 0\}$ is the residual negation of T. This is the most appropriate choice for extracting a strict fuzzy ordering from a given fuzzy ordering L.

From this construction we can infer that the fuzzy relation

$$R_r(x,y) = min(1, max(0, \frac{1}{r}(y-x)))$$

is a T_L-E_r-ordering.

If T_L-E-ordering "L" is strongly complete then it can be proved that a fuzzy relation R defined in (1) is simplified to

Strict fuzzy ordering

where $N_T(x) = sup\{y \in [0,1] | T(x,y) = 0\}$ is the residual negation of T. This is the most appropriate choice for extracting a strict fuzzy ordering from a given fuzzy ordering L.

From this construction we can infer that the fuzzy relation

$$R_r(x,y) = min(1, max(0, \frac{1}{r}(y - x)))$$

is a T_L-E_r-ordering.

If T_L-E-ordering "L" is strongly complete then it can be proved that a fuzzy relation R defined in (1) is simplified to

$$\boxed{R(x,y) = 1 - L(y,x)}$$

Strict fuzzy ordering

and that the following hold -

$$R(x,y) + E(x,y) + R(y,x) = 1 \tag{2}$$

Strict fuzzy ordering

and that the following hold -

$$R(x, y) + E(x, y) + R(y, x) = 1 \tag{2}$$

$$min(R(x, y), R(y, x)) = 0 \tag{3}$$

Fuzzy Ordering-Based rank correlation coefficient

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$.

Fuzzy Ordering-Based rank correlation coefficient

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$. Further assume that E_X and E_Y are two T_L-equivalences,

Fuzzy Ordering-Based rank correlation coefficient

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$. Further assume that E_X and E_Y are two T_L-equivalences, L_X and L_Y a strong complete T_L-E_X-ordering, T_L-E_Y-ordering resp.

Fuzzy Ordering-Based rank correlation coefficient

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$. Further assume that E_X and E_Y are two T_L-equivalences, L_X and L_Y a strong complete T_L-E_X-ordering, T_L-E_Y-ordering resp. Then we can define strict T_L-E_X-ordering on X as $R_X = 1 - L_X(x_2, x_1)$ and a strict T_L-E_Y-ordering on Y as $R_Y = 1 - L_Y(y_2, y_1)$

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$. Further assume that E_X and E_Y are two T_L-equivalences, L_X and L_Y a strong complete T_L-E_X-ordering, T_L-E_Y-ordering resp. Then we can define strict T_L-E_X-ordering on X as $R_X = 1 - L_X(x_2, x_1)$ and a strict T_L-E_Y-ordering on Y as $R_Y = 1 - L_Y(y_2, y_1)$

Given a pair (i, j), we can compute the degree to which (i, j) is **concordant pair** as

$$\tilde{C}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_i, y_j))$$

Fuzzy Ordering-Based rank correlation coefficient

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$. Further assume that E_X and E_Y are two T_L-equivalences, L_X and L_Y a strong complete T_L-E_X-ordering, T_L-E_Y-ordering resp. Then we can define strict T_L-E_X-ordering on X as $R_X = 1 - L_X(x_2, x_1)$ and a strict T_L-E_Y-ordering on Y as $R_Y = 1 - L_Y(y_2, y_1)$

Given a pair (i, j), we can compute the degree to which (i, j) is **concordant pair** as

$$\tilde{C}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_i, y_j))$$

and the degree to which (i, j) is a **discordant pair** as

$$\tilde{D}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_j, y_i))$$

Fuzzy Ordering-Based rank correlation coefficient

Assume that the data is given as (x_i, y_i) with $x_i \in X$ and $y_i \in Y$ for all $i = 1, 2, ..., n$. Further assume that E_X and E_Y are two T_L-equivalences, L_X and L_Y a strong complete T_L-E_X-ordering, T_L-E_Y-ordering resp. Then we can define strict T_L-E_X-ordering on X as $R_X = 1 - L_X(x_2, x_1)$ and a strict T_L-E_Y-ordering on Y as $R_Y = 1 - L_Y(y_2, y_1)$

Given a pair (i, j), we can compute the degree to which (i, j) is **concordant pair** as

$$\tilde{C}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_i, y_j))$$

and the degree to which (i, j) is a **discordant pair** as

$$\tilde{D}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_j, y_i))$$

where $\bar{T}$ is some t-norm to aggregate the relationships of x and y components.

If we adopt the simple sigma count idea to measure the cardinality of a fuzzy set, we can compute the numbers of concordant pairs $\tilde{C}$ and discordant pairs $\tilde{D}$, respectively, as

Fuzzy Ordering-Based rank correlation coefficient

If we adopt the simple sigma count idea to measure the cardinality of a fuzzy set, we can compute the numbers of concordant pairs $\tilde{C}$ and discordant pairs $\tilde{D}$, respectively, as

$$\tilde{C} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{C}(i,j) \qquad \tilde{D} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}(i,j)$$

Fuzzy Ordering-Based rank correlation coefficient

If we adopt the simple sigma count idea to measure the cardinality of a fuzzy set, we can compute the numbers of concordant pairs $\tilde{C}$ and discordant pairs $\tilde{D}$, respectively, as

$$\tilde{C} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{C}(i,j) \qquad \tilde{D} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}(i,j)$$

So the fuzzy Ordering-Based rank correlation measure $\tilde{\gamma}$ as

Fuzzy Ordering-Based rank correlation coefficient

If we adopt the simple sigma count idea to measure the cardinality of a fuzzy set, we can compute the numbers of concordant pairs $\tilde{C}$ and discordant pairs $\tilde{D}$, respectively, as

$$\tilde{C} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{C}(i,j) \qquad \tilde{D} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}(i,j)$$

So the fuzzy Ordering-Based rank correlation measure $\tilde{\gamma}$ as

$$\boxed{\tilde{\gamma} = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}}}$$

Experiment 1

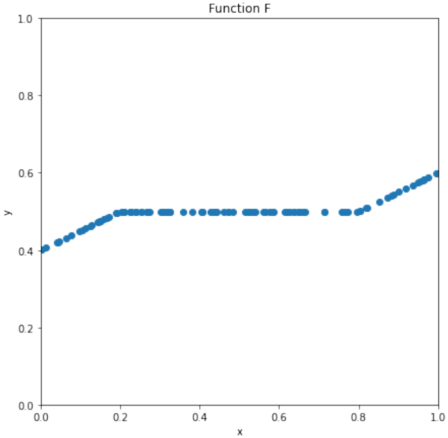
We are given 100 uniformly distributed random values $(x_1, x_2, ..., x_{100})$ from the unit interval.

Experiment 1

We are given 100 uniformly distributed random values $(x_1, x_2, ..., x_{100})$ from the unit interval. The list $(y_1, y_2, ..., y_{100})$ is computed as $y_i = f(x_i)$, where f is a simple, piecewise linear, non-decreasing function that has a relatively large flat area.

Experiment 1

We are given 100 uniformly distributed random values $(x_1, x_2, ..., x_{100})$ from the unit interval. The list $(y_1, y_2, ..., y_{100})$ is computed as $y_i = f(x_i)$, where f is a simple, piecewise linear, non-decreasing function that has a relatively large flat area.



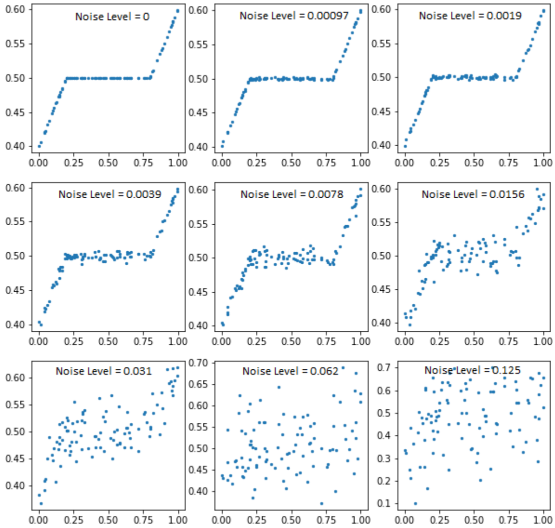


Figure- Different data sets obtained from contaminating a non-decreasing relationship by normally distributed noise with different standard deviations.

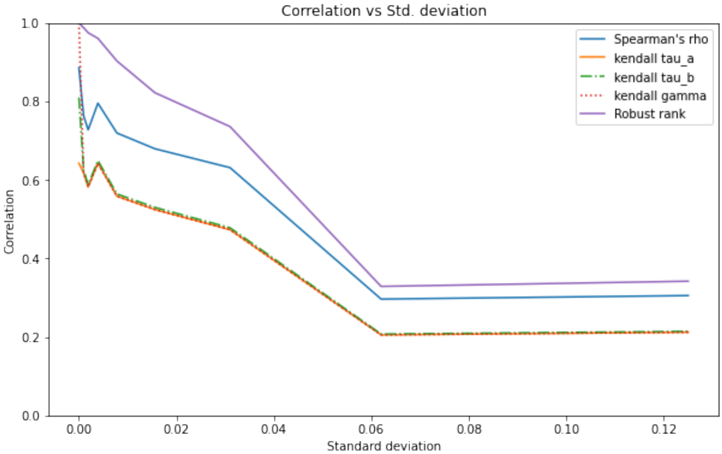


Figure: Results obtained by applying different correlation measures to the data sets

Result (Experiment 1)

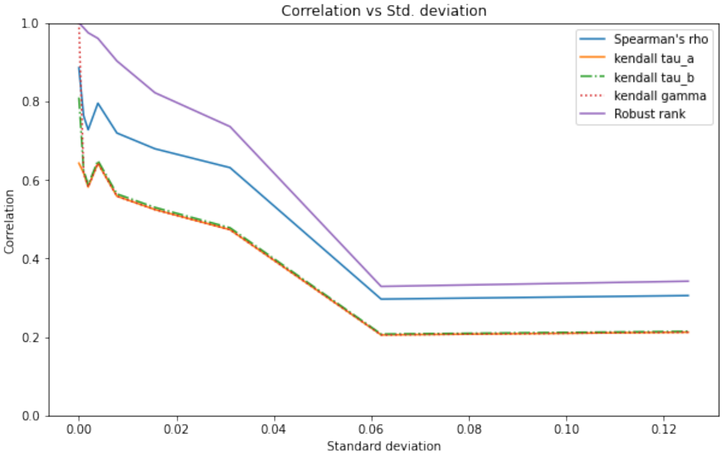


Figure: Results obtained by applying different correlation measures to the data sets

The three lines τ_a, τ_b and γ coincide except for no noise. Lines reveal that these measures react to noise in a **non-robust** way.

Result (Experiment 1)

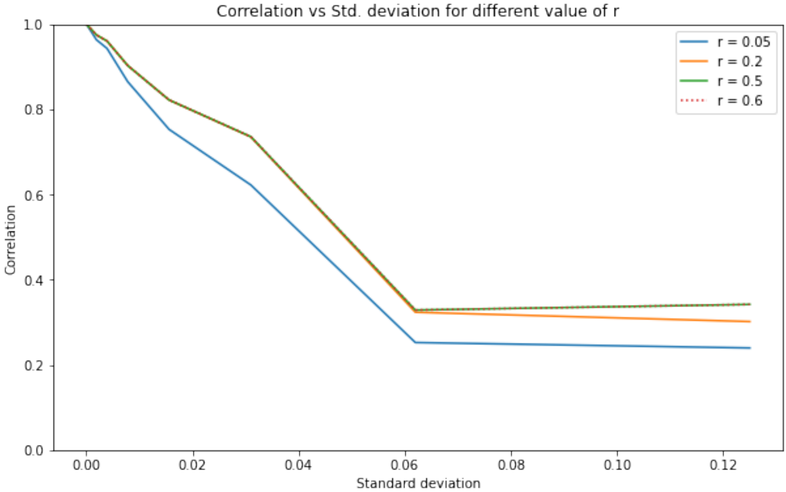


Figure: Results obtained by using different value of r in Robust rank correlation measures to the data sets

Result (Experiment 1)

- All the different variants react to the noise in a more robust way than the crisp measures.

Result (Experiment 1)

- All the different variants react to the noise in a more robust way than the crisp measures.
- The higher r, the more noise is neglected.

Result (Experiment 1)

- All the different variants react to the noise in a more robust way than the crisp measures.
- The higher r, the more noise is neglected.

Note- larger the r, more difficult it is for $\tilde{\gamma}$ to find out whether there are slightly non-monotonic parts in the data.

Experiment 2

This time noise level is fixed $\sigma = 0.01$.

Experiment 2

This time noise level is fixed $\sigma = 0.01$.

Right of $x = 0.5$, we use $f(x) = \frac{x}{2} + \frac{1}{4}$ and, to the left of $x = 0.5$, we linearly interpolate between $(0, q)$.

Experiment 2

This time noise level is fixed $\sigma = 0.01$.

Right of $x = 0.5$, we use $f(x) = \frac{x}{2} + \frac{1}{4}$ and, to the left of $x = 0.5$, we linearly interpolate between $(0, q)$.

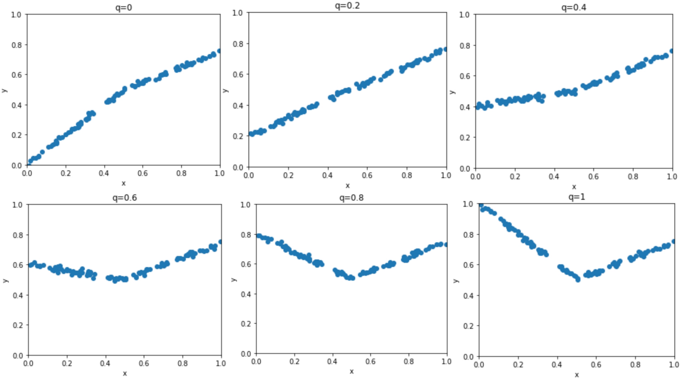


Figure: Noisy data sets that correspond to monotonic ($q \leq 0.5$) and non-monotonic relationships ($q > 0.5$).

Result (Experiment 2)

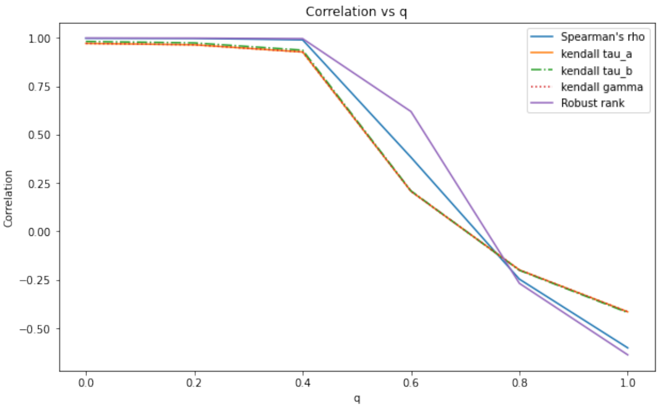


Figure: Results obtained by applying different correlation measures to the data sets.

Result (Experiment 2)

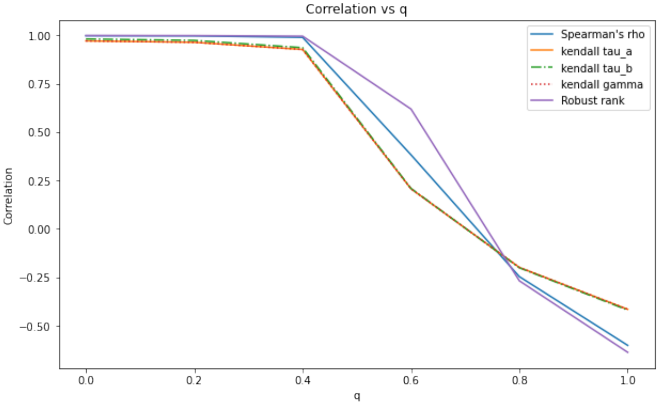


Figure: Results obtained by applying different correlation measures to the data sets.

ρ, τ_a, τ_b and γ again have problems to handle the noise in case of the large proportion of ties that occurs for $q = 0.5$.

Result (Experiment 2)

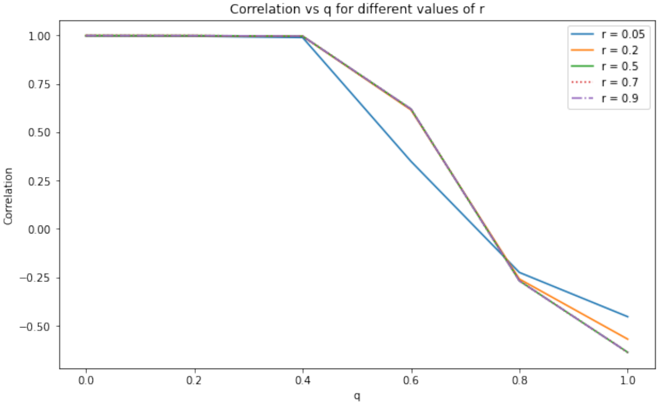


Figure: Results obtained by applying Robust rank correlation measures to the data sets for different values of r.

Result (Experiment 2)

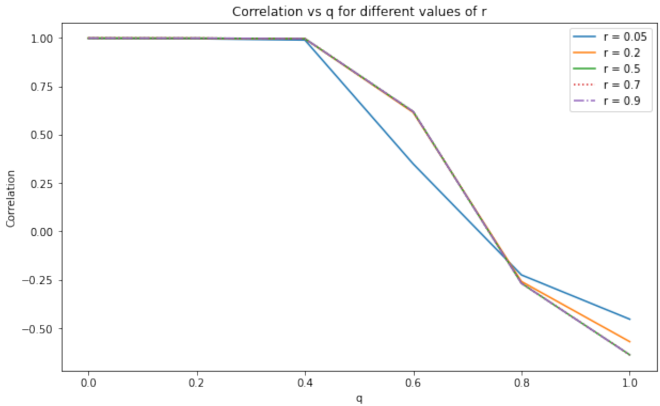


Figure: Results obtained by applying Robust rank correlation measures to the data sets for different values of r.

- All the variants of $\tilde{\gamma}$ show acceptable results for $q \leq 0.5$.

Result (Experiment 2)

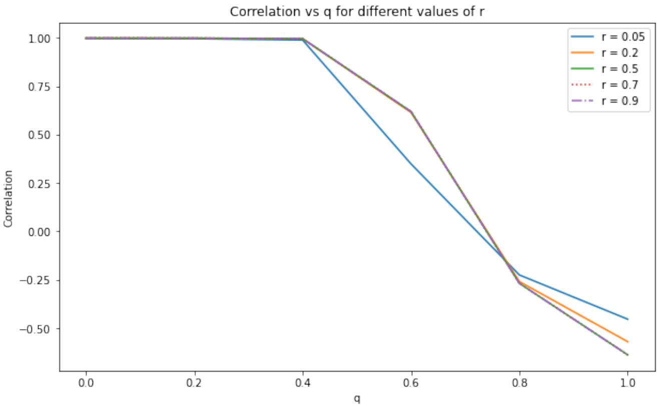


Figure: Results obtained by applying Robust rank correlation measures to the data sets for different values of r.

- All the variants of $\tilde{\gamma}$ show acceptable results for $q \leq 0.5$.
- $\tilde{\gamma}$ yields significantly lower values for $q = 0.6$ in the case $r = 0.05$ which is not in the case of large value of r.

Conclusion

- Robust rank coefficient perform better with noise as compare to other measure.

Conclusion

- Robust rank coefficient perform better with noise as compare to other measure.
- When choosing r, there is a trade-off between robustness (the larger r, the better) and sensitivity (the smaller r, the better).

References

- Robust Rank Correlation Coefficients on the Basis of Fuzzy Orderings: Initial Steps by U. Bodenhofer and F. Klawonn
- The Kendall rank correlation coefficient by H. Abdi. In N. J. Salkind, editor, Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, 2007.