Dependency Measures

Taewan Kim

May 5, 2021

Contents

1 Dependeny Measure Definition **1**
1.1 Pearson's Correlation 1
1.2 Spearman's Correlation 2
1.3 Kendall's Correlation 3
1.4 Distance Correlation 4
1.5 Hoeffding's D Measure 5
1.6 Mutual Information 6
1.7 Maximal Information Coefficient 7
1.8 Total Information Coefficient 7
1.9 Chatterjee's Correlation 8
1.10 Hilbert Schmidt Independence Criterion 9
1.11 Blomqvist's Beta 10

1 Dependeny Measure Definition

1.1 Pearson's Correlation

Pearson's product-moment correlation coefficient or Pearson's correlation is to measure **linear dependeny** of data, as the slope of Y by X in linear regression is calculated by multiplying Pearson's correlation with the ratio of standard deviations.

"Pearson's correlation"

$$\rho_{Pearson}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$where\ \ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}\ \ and\ \ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- If the data has a perfect linear dependence, $\rho_{Pearson}(X,Y) = \pm 1$

- If the data is linearly independent, $\rho_{Pearson}(X, Y) = 0$.
- Lastly, $-1 < \rho_{Pearson}(X, Y) < 1$ if the data has imperfect linear dependence.
- Library: stats, Function Name: cor
  Code Example: stats::cor(dat, method = "pearson")

1.2 Spearman's Correlation

Spearman's rank correlation coefficient or Spearman's correlation is to measure **monotonic relationships** between data (i.e., if values of Y increases, X values should also increases). Spearman's correlation does not require assumptions of linearity between variables, contrast to the Pearson's correlation. For Spearman's Correlation, we should convert the raw values of x_i and y_i to ranks and calculate the differences d_i between the ranks of x_i and y_i. The mathematical definition of Spearman's correlation is as followed.

"Spearman's correlation"

$$\rho_{Spearman}(X, Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

- If the data has a perfect monotonic relationship, $\rho_{Spearman}(X, Y) = \pm 1$
- If the data is monotonically independent, $\rho_{Spearman}(X, Y) = 0$.
- Lastly, $-1 < \rho_{Spearman}(X, Y) < 1$ if the data has imperfect monotonic relationship.
- Library: stats, Function Name: cor
  Code Example: stats::cor(dat, method = "spearman")

1.3 Kendall's Correlation

Kendall τ rank correlation coefficient or Kendall's correlation is also to measure **monotonic relationships** between data, similar to Spearman's correlation.

"Kendall's correlation"

$$\rho_{Kendall}(X,Y) = \frac{\#\ concordant\ pairs - \ \#\ discordant\ pairs}{0.5n(n-1)}$$

- Concordant means if the ranks of both elements agree. (i.e., $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$)
- Discordant means if the ranks of both elements disagree. (i.e., $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$)
- The pair is neither concordant nor discordant if (i.e., $x_i = x_j$ or $y_i = y_j$)

- If the data has a perfect monotonic relationship, $\rho_{Kendall}(X,Y) = \pm 1$
- If the data is monotonically independent, $\rho_{Kendall}(X,Y) = 0$.
- Lastly, $-1 < \rho_{Kendall}(X,Y) < 1$ if the data has imperfect monotonic relationship.
- Library: pcaPP, Function Name: cor.fk
  Code Example: pcaPP::cor.fk(dat)

1.4 Distance Correlation

Distance correlation is to measure **non-linear relationships** between two random variables. As its name implies, the correlation utilizes statistical distance between probability distributions. More specifically, distance correlation is calculated by dividing the distance covariance between X and Y by the product of their distance standard deviations. Mathematical formula comes as followed.

"Distance Correlation"

- Let $||\cdot||$ be the Euclidian distance, $a_{k,l} = ||x_k - x_l||$ and $b_{k,l} = ||y_k - y_l||$
- $\bar{a_k}$ be the kth row mean, $\bar{a_{,l}}$ be the lth column mean and $\bar{a}$ be the grand mean of the distance matrix of X.
- Similarly, $\bar{b_k}$ be the kth row mean, $\bar{b_{,l}}$ be the lth column mean and $\bar{b}$ be the grand mean of the distance matrix of Y.
- $A_{k,l} = a_{k,l} - \bar{a_k} - \bar{a_{,l}} + \bar{a}$ and $B_{k,l} = b_{k,l} - \bar{b_k} - \bar{b_{,l}} + \bar{b}$

$$dCov(X,Y) = \sqrt{\frac{1}{n^2}\sum_{k=1,l=1}^{n} A_{k,l}B_{k,l}}$$

(dVar(X) = dCov(X,X) and dVar(Y) = dCov(Y, Y))

$$dCor(X,Y) = \frac{dCov(Y,Y)}{\sqrt{dVar(X)dVar(Y)}}$$

- The range of dCor is $0 \leq dCor \leq 1$.
- If dCor is 1, the variables have perfect linear relationship(dependence).
- If $0 < dCor < 1$, the variables have imperfect linear dependence.
- if dCor is 0, the variables have indepdence.
- Library: energy, Function Name: dcor2d
  Code Example: energy::dcor2d(x, y)

1.5 Hoeffding's D Measure

Hoeffding's D measure is to test the independence of data by **calculating the distance** between the product of the **marginal distributions** under the null hypothesis and the **empirical bivariate distribution**.

"Hoeffding's D Measure"

- Let R_i and S_i be the rank of x_i and y_i.
- Q_i be the number of points with both x and y values less than the ith point. $Q_i = \sum_{j=1}^{n} \text{ø}(x_j, x_i)\text{ø}(y_j, y_i)$ where $\text{ø}(a, b) = 1$ if $a < b$ and $\text{ø}(a, b) = 0$ otherwise.
- In other words, Q_i is the number of bivariate observations (x_j, y_j) for which $x_j < x_i$ and $y_j < y_i$.

$$D_1 = \sum_{i=1}^{n} Q_i(Q_i - 1), D_2 = \sum_{i=1}^{n} (R_i - 1)(R_i - 2)(S_j - 1)(S_j - 2)$$

$$D_3 = \sum_{i=1}^{n} (R_i - 2)(S_i - 2)Q_i$$

$$D(X, Y) = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

- Differently of the previous coefficients, the positive and negative signs of ρ_n do not have any interpretations, identifying non-monotonic relationships also.
- Library: Hmisc, Function Name: hoeffd
  Code Example: Hmisc::hoeffd(dat)$D

1.6 Mutual Information

Mutual Information is to measure how much one random variable gives information about the other.

"Mutual Information"

$$MI(X,Y) = \int\int f_{X,Y}(x,y)log_2\frac{f_{X,Y}(x,y)}{f_X x f_Y y}\ dxdy \tag{1}$$

$$= \sum\sum p_{X,Y}(x,y)log\frac{p_{X,Y}(x,y)}{P_X(x)P_Y(y)} \tag{2}$$

- We use (1) if X, Y are continuous variables.
- We use (2) if X, Y are discrete variables.

- Mutual information only has postivie values.
- High mutual information indicates a large reduction in uncertainty.
- Low mutual information indicates a small reduction in uncertainty.
- Zero mutual information between two random variables indicates that the variables are independent.
- Library: entropy, Function Name: discretize2d
  Code Example
  discretized <- entropy::discretize2d(x,y, numBins1 = 20, numBins2 = 20)
  entropy::mi.empirical (discretized)
- Since mutual information does not have a bounded range, i.e., $MI = [0, \inf)$, we use normalized mutual information for our research.
  Normalized mutual information is defined as $NMI(X,Y) = \frac{2*I(X;Y)}{[H(X)+H(Y)]}$

1.7 Maximal Information Coefficient

Maximal Information Coefficient is based on the idea that a grid can be drawn on the scatterplot of the two variables if there is a relationship between two random variables. The characteristic matrix M is defined as $M = (m_{a,b})$, where $(m_{a,b})$ is the highest normalized MI achieved by an a-by-b grid. The statistic MIC is the maximum value in M.

- Library: minerva, Function Name: mine
  Code Example: minerva::mine(dat)$MIC

1.8 Total Information Coefficient

Total Information Coefficient (TIC) is designed to be used as a test statistic for testing a null hypothesis of independence. More specific, TIC tries to prevent positive bias in case of statistical independence from MIC, which may neglects some useful information by taking the maximal value of characteristic matrix. There are two types of TIC, given by Reshef el al. (2015): the characteristic matrix (TIC) and the equicharacteristic matrix (TIC_e).

"Total Information Coefficient"

$\widehat{M}$ denotes a sample characteristic matrix, $\widehat{M}(D)_{k,l} = \frac{I^*(D,k,l)}{logmin\{k,l\}}$

$[\widehat{M}]$ denotes a sample equicharacteristic matrix, $[\widehat{M}](D)_{k,l} = \frac{I^{[*]}(D,k,l)}{logmin\{k,l\}}$

$D \in \mathbb{R}^2$, and $B : \mathbb{Z}^+ \to \mathbb{Z}^+$

$$TIC_B(D) = \sum_{kl \leq B(n)} \widehat{M}(D)_{k,l} \tag{1}$$

$$TIC_{e,B}(D) = \sum_{kl \leq B(n)} [\widehat{M}](D)_{k,l} \tag{2}$$

- (1) is the characteristic matrix (TIC)
- (2) is the equicharacteristic matrix (TIC_e)

- If (X,Y) are a pair of jointly distributed random varialbes and statistically independent, $M(X, Y) \equiv [M](X, Y) \equiv 0$
  If not, there exists some $a > 0$ and some integer $l_0 \geq 2$ such that

  $$M(X, Y)_{k,l}[M](X, Y)_{k,l} \geq \frac{a}{logmin\{k, l\}}$$

  either for all $k \geq l \geq l_0$, or for all $l \geq k \geq l_0$

1.9 Chatterjee's Correlation

Chatterjee's correlation is to measure the degree of dependence between the variables. In order to calculate the correlation, we should rearrange the data as $(X_{(1)}, Y_{(1)}), ..., (X_{(n)}, Y_{(n)})$ such that $X_1 \leq \cdots \leq X_n$. Let r_i be the rank of $Y_{(i)}$, which is the number of j such that $Y_{(j)} \geq Y_{(i)}$.

"Chatterjee's Correlation"

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

$$\xi_n(X,Y) = 1 - \frac{n]\sum_{i=1}^{n-1} |r_{i+1} - r_i}{2\sum_{i=1}^{n} l_i(n - l_i)} \tag{4}$$

- We use (1) if the X_i's and the Y_i's have no ties.
- We use (2) if there are ties in X_i's or Y_i's

- ξ_n is not symmetric in X and Y, unlike most coefficients.
- $0 \leq \xi_n \leq 1$. If $\xi(X,Y) = 0$, X and Y are independent. Or, if $\xi(X,Y) =$, Y is a measurable function of X.
- $\xi_n(X,Y)$ remains unchanged if we apply strictly increasing transformations to X and Y, because it is based on ranks.
- Library: XICOR, Function Name: calculateXI
  Code Example: XICOR::calculateXI(x, y)

1.10 Hilbert Schmidt Independence Criterion

Hilbert Schmidt Independence Criterion, HSIC, is based on a distance between the join distribution $f_{X,Y}$ and the product of its marginal $f_X f_Y$ to test whether the variables are independent. This is done throught the kernel embeddings of probability measures into reproducing kernal Hilbert spaces. HSIC utilizes the maximum mean discrepancy(MMD) between two probability measures, the norm-induced metric in the RKHS. After all, HSIC measure is then defined as the MMD between the join and the product distributions.

"Hilbert Schmidt Independence Criterion"

$$
\begin{aligned}
H =& E\{k(X,X')l(Y,Y')\} + E\{k(X,X')\}E\{l(Y,Y')\} \\
& - 2E\{E\{k(X,X')|X\}E\{l(Y,Y')|Y\}\}
\end{aligned}
$$

where (X, Y) and (X', Y') are i.i.d of P_{XY}, $k(X, X')$ and $l(Y, Y')$ are kernel functions.

- Library: dHSIC, Function Name: dhsic
  Code Example: dHSIC::dhsic(x, y)

1.11 Blomqvist's Beta

Blomqvist's Beta or medial correlation coefficient is to measure dependency between two variables by constructing two-way contingency table with the medians of each margin as cutting points.

- Let n_1 be the number of data in the first or third quadrant of the table.
- Let n_2 be the number of data belonging to the second or fourth quadrant.

"Blomqvist's Beta"

$$\beta_n = \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1 \tag{5}$$

$$\beta = P\{(X - \widetilde{x})(Y - \widetilde{y}) > 0\} - P\{(X - \widetilde{x})(Y - \widetilde{y}) < 0\} \tag{6}$$

where $\widetilde{x}$ and $\widetilde{y}$ are the median of X and Y, respectively.

- (3) is the original definition of Blomqvist's Beta, based on the contingency table.
- We use (4) for the pair of continuous random variables X and Y and the population version of $\widehat{\beta_n}$ in (3).
- Library: VineCopula, Function Name: BetaMatrix
  Code Example: VineCopula::BetaMatrix(dat)$dHSIC

References

https://wisostat.uni-koeln.de/fileadmin/sites/statistik/pdf_publikationen/SchmidSchmidtBlomqvistsBeta.pdf.