Journal of the Korean Data & Information Science Society
2015, **26**(4), 1017–1025

http://dx.doi.org/10.7465/jkdi.2015.26.4.1017
한국데이터정보과학회지

An application of mutual information in mathematical statistics education[†]

Seongbaek Yi[1] · Dae-Heung Jang[2]

[12]Department of Statistics, Pukyong National University

Received 19 June 2015, revised 16 July 2015, accepted 20 July 2015

Abstract

In mathematical statistics education, we can use mutual information as a tool for evaluating the degree of dependency between two random variables. The ordinary correlation coefficient provides information only on linear dependency, not on nonlinear relationship between two random variables if any. In this paper as a measure of the degree of dependency between random variables, we suggest the use of symmetric uncertainty and λ which are defined in terms of mutual information. They can be also considered as generalized correlation coefficients for both linear and non-linear dependence of random variables.

Keywords: Independence, mutual information, symmetric uncertainty.

1. Introduction

A lot of attention has been given to the mutual information as interdisciplinary subjects - see Wu *et al.* (2009), Vretos *et al.* (2011), Gomez-Verdejo *et al.* (2012) and Zeng *et al.* (2012). In probability and information theory, mutual information is defined as a measure of the amount of information that one random variable contains about another random variable. When two random variables X and Y have a joint probability distribution $f(x,y)$ and marginal probability functions $f_X(x)$ and $f_Y(y)$, respectively, Cover and Thomas (1991) defines the mutual information $\mathrm{MI}(X,Y)$ as the relative entropy between the joint distribution $f(x,y)$ and the product of marginal distributions $f_X(x)f_Y(y)$, i.e.,

$$\mathrm{MI}(X,Y) = \begin{cases} \iint f(x,y)\log\left(\frac{f(x,y)}{f_X(x)f_Y(y)}\right)dxdy & \text{if continuous} \\ \sum_x\sum_y f(x,y)\log\left(\frac{f(x,y)}{f_X(x)f_Y(y)}\right) & \text{if discrete.} \end{cases} \tag{1.1}$$

Since the entropy is a measure of uncertainty of a random variable, the mutual information $\mathrm{MI}(X,Y)$ can be considered as the reduction in the uncertainty of one random variable by the knowledge of the other random variable.

[†] This work was supported by a Research Grant of Pukyong National University (CD-2014-0623).
[1] Corresponding author: Professor, Department of Statistics, Pukyong National University, Busan 608-747, Korea. E-mail: sbyi0108@gmail.com.
[2] Professor, Department of Statistics, Pukyong National University, Busan 608-747, Korea.

Let the joint entropy of two random variables X and Y with joint distribution $f(x, y)$ be defined as $H(X, Y) = -E[\log f(X, Y)]$. The mutual information can be rewritten as

$$\mathrm{MI}(X, Y) = H(X) + H(Y) - H(X, Y), \tag{1.2}$$

where $H(X)$ and $H(Y)$ are the marginal entropies of X and Y respectively. The mutual information has nonnegative value, i.e., $\mathrm{MI}(X, Y) \geq 0$. If $\mathrm{MI}(X, Y) = 0$, two random variables are independent and otherwise they are dependent. There have been proposed several normalized variations of mutual information for various needs. We consider the use of the symmetric uncertainty (SU) by Witten and Frank (2005), and the global correlation coefficient λ by Darbellay (1998), defined respectively by

$$\mathrm{SU}(X, Y) = 2\left(\frac{\mathrm{MI}(X, Y)}{H(X) + H(Y)}\right) \quad \text{and} \quad \lambda = \sqrt{1 - e^{-2\mathrm{MI}(X,Y)}}. \tag{1.3}$$

Since $\mathrm{MI}(X, Y) = H(X) + H(Y) - H(X, Y)$, we know that they are normalized, i.e.,

$$0 \leq SU(X, Y) \leq 1 \quad \text{and} \quad 0 \leq \lambda \leq 1. \tag{1.4}$$

When two random variables are independent, SU (or λ) is equal to zero, while the degree of dependency becomes high as they get close to one. With SU (or λ) equal to one, we can predict exactly one random variable from the other. This motivates the use of symmetric uncertainty or λ as generalized correlation coefficients for evaluating the degree of dependence which can not be guessed with ordinary correlation coefficients.

In section 2 we apply the measures for various quantitative variables of both continuous and discrete random variables as well. Section 3 has the same computations for the simulated sample data, section 4 has for categorical data and the final section summarizes the study.

2. Quantitative variables

Let $f(x, y)$ denote the joint probability density function of random variables X and Y, and $f_X(x)$ and $f_Y(y)$ the marginal probability density functions of X and Y respectively. The random variables X and Y with respective supports $\mathcal{X}$ and $\mathcal{Y}$ are said to be independent if

$$f(x, y) = f_X(x) f_Y(y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2.1}$$

If the equality does not hold, two random variables are said to be dependent. The definition, however, does not yield any information about the degree of dependence when they are dependent. It is common to use correlation coefficient to measure the relationship between two random variables. But the correlation coefficient tells only the degree of *linear* dependency between them.

When two variables are independent, the resulting correlation coefficient is zero. But the converse is not true because the correlation coefficient detects only linear dependency between two variables. Suppose random variable X has a distribution symmetric about zero, and consider $Y = X^2$. The random variable Y is completely determined by the random variable X, thus X and Y are perfectly dependent. But the correlation coefficient of X and Y is zero.

The following examples show the degree of dependency for dependent random variables using the measure symmetric uncertainty SU or λ.

Example 2.1 We consider two random variables X and Y with the joint probability density function

$$f(x,y) = \frac{1}{2}e^{-y}, \quad -y \le x \le y, \quad 0 \le y \le \infty.$$

The marginal probability density functions of X and Y are $f_X(x) = \frac{1}{2}e^{-|x|}$ and $f_Y(y) = ye^{-y}$, respectively. The 3D surface plot and the contour plot of the joint probability density function are shown in Figure 2.1. The correlation coefficient of X and Y is zero, which implies no linear relationship between the two variables. But the fact the equation (2.1) does not hold for two variables implies they are dependent. Using mutual information we can guess how much they are dependent. We have $\mathrm{MI}(X,Y) = 0.5772$, $\mathrm{SU}(X,Y) = 0.7319$, and $\lambda = 0.8275$, implying strong nonlinear dependency between them.

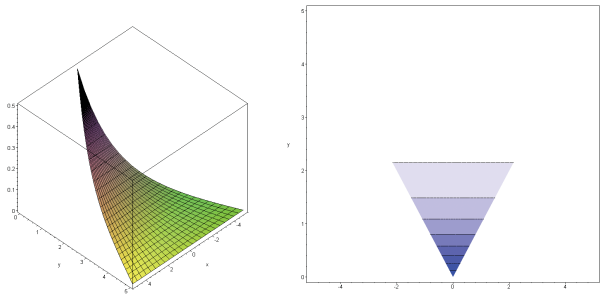


Figure 2.1 Surface and contour plot of the joint probability density function in example 2.1

Example 2.2 Suppose we have random variables X and Y with the joint probability density function $f(x,y)$, defined by

$$f(x,y) = (1+\alpha) - 2\alpha(x+y-2xy),$$

where $|\alpha| \le 1, 0 \le x \le 1, 0 \le y \le 1$. The marginal probability density functions of X and Y are $f_X(x) = 1$ and $f_Y(y) = 1$, respectively. When $\alpha = 0$, the joint probability density function is the same as bivariate uniform density function.

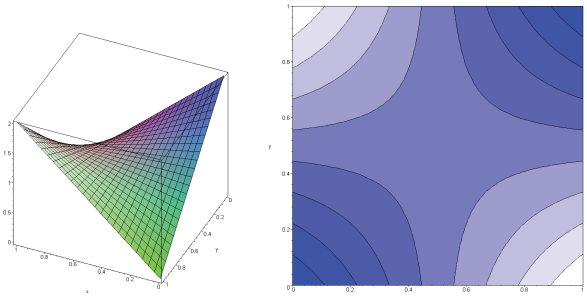


Figure 2.2 Surface and contour plot of joint pdf in example 2.2 with $\alpha = 1$

In Figure 2.2 we show the 3D surface plot and the contour plot of the joint probability density function corresponding to $\alpha = 1$. The correlation coefficient ρ is $\alpha/3$ and the mutual information is

$$MI(X,Y) = \frac{a\alpha^2 + b\alpha + c}{8\alpha},$$

where

$$\begin{aligned} a &= \log(1+\alpha) - \log(1-\alpha), \\ b &= 4\log(1+\alpha) + 4\log(1-\alpha) - 10, \\ c &= 3\log(1+\alpha) - 3\log(1-\alpha) - 2\text{dilog}(1+\alpha) + 2\text{dilog}(1+\alpha), \\ \text{dilog}(x) &= \int_1^x \log(t)/(1-t)dt. \end{aligned}$$

The plots of correlation coefficient, mutual information, and λ for various α values are shown in Figure 2.3. We have an interesting feature that $|\rho| \approx \lambda$. When $\alpha = 0$, all the values of correlation coefficient, mutual information, and λ are zero. Also random variables X and Y are independent by the definition of equation (2.1).

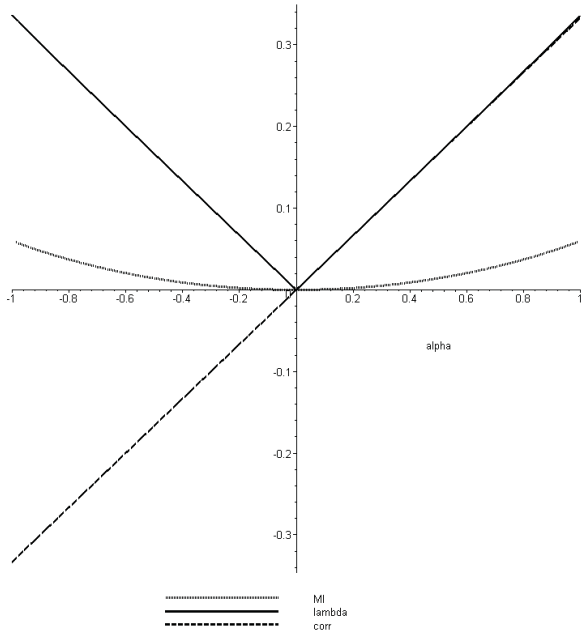


Figure 2.3 Plots of correlation, mutual information, and λ with varying α in example 2.3

Example 2.3 Let X and Y have the joint probability mass function described by the following table:

$Y \backslash X$	-1	0	1	total
0	0	1/3	0	1/3
1	1/3	0	1/3	2/3
total	1/3	1/3	1/3	1

The correlation coefficient of X and Y is zero, which implies that there is no linear relationship between the two variables. However they are dependent since the equation (2.1) does not hold. As a measure of the degree of dependence, we have $\mathrm{MI}(X,Y) = 0.6365$, $\mathrm{SU}(X,Y) = 0.7337$ and $\lambda = 0.8485$, implying strong nonlinear dependency between them.

Example 2.4 Consider an experiment of rolling a die and tossing a coin. Let X denote the appeared number of the die. We define an indicator variable whose value is one for head and zero for tail when tossing the coin. Let the random variable Y be the sum of the number of the die and the value of the indicator variable. Then we have the following joint probability mass function of X and Y:

$X\backslash Y$	1	2	3	4	5	6	7	total
1	1/12	1/12	0	0	0	0	0	1/6
2	0	1/12	1/12	0	0	0	0	1/6
3	0	0	1/12	1/12	0	0	0	1/6
4	0	0	0	1/12	1/12	0	0	1/6
5	0	0	0	0	1/12	1/12	0	1/6
6	0	0	0	0	0	1/12	1/12	1/6
total	1/12	1/6	1/6	1/6	1/6	1/6	1/12	1

The random variables X and Y have correlation coefficient 0.9597, implying strong positive correlation. We also have $\mathrm{MI}(X,Y) = 1.2141$, $\mathrm{SU}(X,Y) = 0.6565$, and $\lambda = 0.9549$, specifically ρ almost equal to λ.

Example 2.5 Let the random variables X and Y follow bivariate normal distribution with respective variance σ_X^2 , σ_Y^2 and correlation coefficient ρ. Then

$$H(X) = \frac{1}{2}\log(2\pi) + \log\sigma_X + \frac{1}{2},$$

$$H(Y) = \frac{1}{2}\log(2\pi) + \log\sigma_Y + \frac{1}{2},$$

$$\mathrm{MI}(X,Y) = -\frac{1}{2}\log(1-\rho^2).$$

Therefore we have

$$\mathrm{SU}(X,Y) = \frac{-\log\sqrt{1-\rho^2}}{\log(2\pi e) + \log(\sigma_X\sigma_Y)}, \quad \text{and} \quad \lambda = |\rho|.$$

Example 2.6 Block and Basu (1974) proposed an absolutely continuous bivariate exponential distribution given by

$$f(x,y) = \begin{cases} \frac{\mu_1\mu(\mu_2+\mu_{12})}{\mu_1\mu_2}\exp\{-\mu_1 x - (\mu_2+\mu_{12})y\} & \text{if } x < y \\ \frac{\mu_2\mu(\mu_1+\mu_{12})}{\mu_1\mu_2}\exp\{-(\mu_1+\mu_{12})x - \mu_2 y\} & \text{if } x > y \end{cases} \tag{2.2}$$

where $x > 0, y > 0$ and μ_1, μ_2, μ_{12} and μ are parameters satisfying $\mu_1 > 0, \mu_2 > 0, \mu_{12} > 0$ and $\mu = \mu_1 + \mu_2 + \mu_{12}$. The distributions of X and Y are independent if and only if $\mu_{12} = 0$,

and if $\mu_1 = \mu_2$, the marginal distributions of X and Y are identical. The individual entropies and mutual information are given respectively by

$$\begin{aligned}
H(X) &= -\ln\mu - \ln\left(\frac{\mu_2+\mu_{12}}{\mu_1+\mu_2}\right) + (\mu_1+\mu_2)\left[\frac{1}{\mu_1+\mu_{12}} + \frac{\mu_2\mu_{12}}{\mu(\mu_1+\mu_2)(\mu_1+\mu_{12})}\right] \\
&\quad + \sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\mu_{12}}{\mu_1+\mu_{12}}\right)^k M(-k\mu_2), \\
H(Y) &= -\ln\mu - \ln\left(\frac{\mu_1+\mu_{12}}{\mu_1+\mu_2}\right) + (\mu_1+\mu_2)\left[\frac{1}{\mu_2+\mu_{12}} + \frac{\mu_1\mu_{12}}{\mu(\mu_1+\mu_2)(\mu_2+\mu_{12})}\right] \\
&\quad + \sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\mu_{12}}{\mu_2+\mu_{12}}\right)^k M(-k\mu_1), \\
\mathrm{MI}(X,Y) &= \frac{\mu\ln a + \mu_2\ln b}{\mu_1+\mu_2} - \ln\mu - \ln\left(\frac{\mu_1+\mu_{12}}{\mu_1+\mu_2}\right) - \ln\left(\frac{\mu_2+\mu_{12}}{\mu_1+\mu_2}\right) + \frac{\mu_{12}}{\lambda} \\
&\quad - \sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\mu_{12}}{\mu_2+\mu_{12}}\right)^k M(-k\mu_1) - \sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\mu_{12}}{\mu_1+\mu_{12}}\right)^k M(-k\mu_2),
\end{aligned}$$

where

$$M(t) = E(e^{tX}) = \frac{1}{\mu_1+\mu_{12}}\cdot\frac{\mu}{\mu-t}\left[\mu_2 + \frac{\mu_1(\mu_2+\mu_{12})}{\mu_2+\mu_{12}-t}\right].$$

If $\mu_1 = \mu_2 = \mu_0$ and $\mu_{12} = \delta\mu_0$, then

$$\mathrm{MI}(X,Y) = -\ln\{(1+\delta)(1+\delta/2)\} + \frac{\delta}{\delta+2} + M_0(\delta),$$

where

$$M_0(\delta) = 2\sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{\delta}{1+\delta}\right)^k \frac{(2+\delta)(2+2\delta-k)}{(1+\delta)(2+\delta-k)(1+\delta-k)}.$$

It is evident that $\mathrm{MI}(X,Y)$ depends only on δ and is equal to zero if $\delta = 0$ and $\mathrm{MI}(X,Y)$ increases as δ gets increased. And the resulting $\mathrm{SU}(X,Y)$ is computed by the definition.

Example 2.7 Suppose we have generalized Gumbel's bivariate logistic (GGBL) distribution, which was first introduced by Gumbel (1961). The joint probability density function of GGBL distribution is given by

$$f(x,y) = \frac{m(m+1)e^{-x-y}}{(1+e^{-x}+e^{-y})^{m+2}}$$

where $-\infty < x, y < \infty$ and $m > 0$. We have the individual entropies and MI as follows:

$$H(X) = H(Y) = -\ln m + \Psi(m) + c + \frac{m+1}{m},$$

$$\mathrm{MI}(X,Y) = \ln(m+1) - \ln m + \frac{1}{m+1},$$

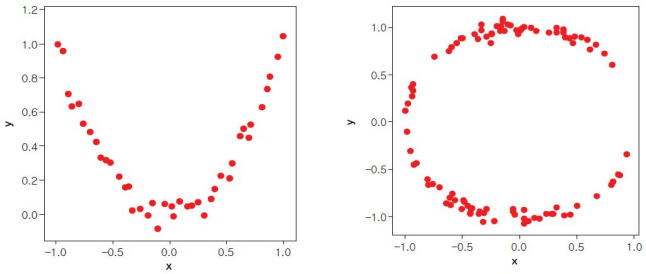
where $\Psi(m) = \Gamma'(m)/\Gamma(m)$, the digamma function and $c = -\Psi(1)$, the Euler's constant. We know that the mutual information decreases monotonically to zero as m increases to ∞.

3. Sample correlation coefficient

Given n pairs of observations $(x_i, y_i), i = 1, 2 \cdots, n$, we use sample correlation coefficient as an estimate of the population correlation coefficient. The sample correlation coefficient gives information only about linearity between two variables, nothing about nonlinearity. We can, however, use symmetric uncertainty or λ using mutual information to guess the amount of dependency in case of nonlinear relationship.

In computing mutual information it is required to know the joint pdf $f(x, y)$, and the marginal pdfs $f_X(x)$ and $f_Y(y)$. Unless we have any idea of their specific forms, we estimate the functions with popular two approaches - bivariate histogram and bivariate kernel estimator. There are two methods for the bivariate histogram approaches such as equidistant cells and equiprobable cells. For more on estimation of mutual information, see Darbellay (1999), Darbellay and Vajda (1999), Harrold *et al.* (2001), Chelikani *et al.* (2003), Kraskov *et al.* (2004) and Zhou *et al.* (2005). The following example shows the MI computation using the estimate of bivariate histogram.

Example 3.1 The correlation coefficients of the data with two scatter plots (a) and (b) at Figure 3.1 are 0.029 and 0.077 respectively. They are close to zero, and have no linear relationship between two variables. However the plot (a) shows the curve $y = x^2$ and the plot (b) the circle $x^2 + y^2 = 1$. The nonlinear relationship in the plots is not uncovered by the sample correlation coefficients. Using the bivariate histogram estimates for the two data sets, we have 1.279 and 0.631 respectively as estimates of the mutual information. They yield 0.961 and 0.847 as the values of λ, respectively, for the two scatter plots, which implies the existence of strong nonlinear relationship.



(a) (b)

Figure 3.1 Scatter plot for non-linear relationship in example 3.1

4. Categorical variables

When we consider categorical variables, we get an estimate of mutual information using the contingency table as follows:

$$\hat{\mathrm{MI}}(X, Y) = \sum_i \sum_j \frac{n_{ij}}{n} \log\left(\frac{n_{ij} n}{n_{i+} n_{+j}}\right), \tag{4.1}$$

where n is the total number of observations, n_{ij} the number of individuals in the cell of the ith category of the variable X and the jth category of the variable Y, n_{i+} the number of individuals among the n sampled falling into the category i of variable X, and n_{+j} the number of individuals among the n sampled falling into the category j of variable Y.

As measures of association between two classification variables, we mainly use statistics such as chi-square test statistic, likelihood chi-square test statistic, and Mantel-Haenszel test statistic. Additionally symmetric uncertainty and λ can also be used.

Example 4.1 Table 4.1 is a two-way table summarizing the number of persons convicted of drunkenness in two London courts during the first six months of 1970 (Hand *et al.*, 1994).

Table 4.1 The number of persons convicted of drunkenness in two London courts

Gender \ Age	0-29	30-39	40-49	50-59	≥ 60	total
Male	185	207	260	180	71	903
Female	4	13	10	7	10	44
total	189	220	270	187	81	947

Table 4.2 shows some of test results for the data in Table 4.1. The p-values of each test procedure tell significant association between two variables. The value of λ by mutual information agrees with the results, giving information about the degree of dependency.

Table 4.2 Test results for the data in example 4.1

Statistics	DF	Value	P-value
Chi-Square	4	15.2461	0.0042
Likelihood Ratio Chi-Square	4	12.6670	0.0130
Mantel-Haenszel Chi-Square	1	4.8961	0.0269
Phi Coefficient		0.1269	
Contingency Coefficient		0.1259	
Cramer's V		0.1269	
Mutual Information		0.0067	
λ		0.1153	

5. Concluding Remarks

The ordinary correlation coefficient provides information only about linear dependency, not about nonlinear relationship between two random variables if any. In this paper as a measure for evaluating the degree of dependency between variables, we have suggested symmetric uncertainty or λ using mutual information. They are considered as generalized correlation coefficients for both linear and non-linear dependence between two random variables.

References

Block, H. B. and Basu, A. P. (1974). A continuous bivariate exponential extension. *Journal of the American Statistical Association*, **69**, 1031-1037.

Chelikani, S., Purushothaman, K. and Duncan, J. S. (2003). Support vector machine density estimator as a generalized Parzen windows estimator for mutual information based image registration. *Lecture Notes in Computer Science*, **2879**, 854-861.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*, Wiley, New York.

Darbellay, G. A. (1998). *An adaptive histogram estimator for the estimator for the mutual information*, Research Report no. 1936, UTIA, Academy of Science, Prague.

Darbellay, G. A. (1999). An estimator of the mutual information based on a criterion for independence. *Computational Statistics and Data Analysis*, **32**, 1-17.

Darbellay, G. A. and Vajda, I. (1999). Estimation of the information by an adaptive partition of the observation space. *IEEE Transactions on Information Theory*, **45**, 1315-1321.

Gomez-Verdejo, V., Martinez-Ramon, M., Florensa-Vila, J. and Oliviero, A. (2012). Analysis of fMRI time series with mutual information. *Medical Image Analysis*, **16**, 451-458.

Gumbel, E. J. (1961). Bivariate logistic distribution. *Journal of the American Statistical Association*, **56**, 335-349.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994). *A handbook of small data sets*, Chapman and Hall, London.

Harrold, T. I., Sharma, A. and Sheather, S. (2001). Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. *Stochastic Environmental Research and Risk Assessment*, **15**, 310-324.

Kraskov, A., Stogbauer, H. and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, **69**, 066138/1-066138/16.

Vretos, N., Solachidis, V., and Pitas, I. (2011). A mutual information based face clustering algorithm for movie content analysis. *Image and Vision Computing*, **29**, 693-705.

Witten, I. H. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann Publishers Inc., San Francisco.

Wu, E. H., Philip, L. H. and Li, W. K. (2009). A smoothed bootstrap test for independence based on mutual information. *Computational Statistics and Data Analysis*, **53**, 2524-2536.

Zhou, G., Yang, L., Su, J. and Ji, D. (2005). Mutual information independence model using kernel density estimation for segmenting and labeling sequential data. *Lecture Notes in Computer Science*, **3406**, 155-166.

Zeng, J., Xie, L., Kruger, U. and Gao, C. (2012). A non-Gaussian regression algorithm based on mutual information maximization. *Chemometrics and Intelligent Laboratory Systems*, **111**, 1-19.