Journal of Selçuk University Natural and Applied Science
Online ISSN: 2147-3781
www.josunas.org

Do some relative entropy measures coincide in determining correlations or associations for metric data?

Atif EVREN[a*] and Gökhan DINCER[b]

[a] *Yildiz Tech. Univ., Faculty of Science & Literature, Dep. of Statistics, ,Davutpasa, Istanbul, TURKEY*
[b] *Yildiz Tech. Univ., Faculty of Science & Literature, Dep. of Statistics ,Davutpasa, Istanbul, TURKEY*

Abstract

Entropy is a measure of uncertainty of a statistical experiment or the measure of information provided by experimentation. Several measures of entropy are used in uncertainty considerations for nominal, ordinal (as well as metric) data and specifically in qualitative variation calculations. Besides, relative entropy concepts (e.g. mutual information, etc.) are used in goodness of fit tests or in checking the adequacy of any statistical model in general. In particular, relative entropy measures are used in correlation or association estimations. In this study, based on a specific definition of mutual information, we use some different relative entropy measures. Then we compare these measures under three different situations by some applications.

Keywords: Association measures based on entropy, mutual information, relative entropy

Introduction

Whenever the scale of data is nominal or ordinal, classic correlation measures can not be applied. In such situations association measures based on frequencies are used. Among these association measures, mutual information has gained popularity recently. Although there has already been a literature on mutual information (relative entropy) and its statistical applications, there is no unique definition of mutual information as a result of different parametrizations or different entropy definitions. In general what is understood by mutual information is the difference between the entropy of a marginal probability distribution and the average conditional entropy. Here it should be emphasized that the entropies are often calculated by Shannon's methodology although there are alternative ways of modeling uncertainty and entropy.

In this study, we compute mutual information statistics alternatively (i.e. by using Rényi and Tsallis formulations) for some hypothetical bivariate distributions, and

try to show that, due to different parametrizations, some misleading results may be obtained.

1. Entropy and information theory

The concepts of entropy and uncertainty are closely associated with each other. If the test result is unpredictable then the amount of entropy is great. So in some sense, the degree of uncertainty and the information one can get from statistical experimentation are related to each other.

Although Planck wrote down his equation in 1872, L. Boltzman is known as the first scientist who emphasized the probabilistic meaning of the thermodynamical entropy. Planck and Boltzman showed that the entropy of a physical system can be considered as a measure of disorder in the system. In case of a physical system having many degrees of freedom (e.g. a perfect gas), the number measuring the disorder of the system measures also the uncertainty concerning the states of the individual particles" (Rényi, 2007a).

2. Shannon, Rényi and Tsallis entropies

Although Shannon entropy is the most commonly used uncertainty measure in practice, Rényi entropy which is used in biological diversity calculations and Tsallis entropy as the general case of Simpson's diversity index , as analytical tools, have been used often by numerous scientists for some years. One can refer to Pardo(2006), Esteban and Morales(1995), Ullah(1996) and Cahill(2010) for a collective exposition of different entropy measures. For elementary concepts and statistical applications of Shannon entropy, Reza(1994), Rényi(2007a and 2007b), Golan(2006) and Cover and Thomas(2006) should be emphasized.

In information theory, Shannon entropy is a special case of Rényi entropy. As $\alpha \to 1$, the limit of Rényi entropy is Shannon entropy. Shannon and Wieder have suggested the following measure of uncertainty or entropy associated with the sample space of a complete finite scheme.

$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

* Corresponding author. Tel.:+90212 383 44 35; fax: +902123834415; e-mail: aevren2006@gmail.com.

Here we assume that the discrete random variable X takes on $x_1, x_2, \dots, x_n$ values with respective probabilities $p_1, p_2, \dots, p_n$. For continuous distributions, summation sign in (1) is simply replaced by integration sign. For two discrete variables X and Y having joint probability function $P_{ij} = P(X_i = x_i; Y_j = y_j)$ for i=1,2,...,n and j=1,2,...,m, we have

$$H = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} log p_{ij} \quad (2)$$

As can be seen, extensions to continuous cases or multivariate cases are straightforward.

Rényi entropy which was named after Alfréd Rényi is defined as

$$H_R^{\alpha}(\mathrm{x}) = \frac{1}{1-\alpha}\log(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P_i}^{\alpha}) \quad (3)$$

when α is a positive constant.

Tsallis entropy is defined as

$$H_T^{\alpha}(\mathrm{x}) = \frac{1-\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p_i}^{\alpha}}{\alpha-1}, \quad \alpha > 0, \alpha \neq 1 \quad (4)$$

It should also be noted that for univariate distributions, Rényi and Tsallis entropies are associated each other so that the following equation holds;

$$H_R^{\alpha}(\mathrm{x}) = \frac{\log(1-(\alpha-1)\, H_T^{\alpha}(x))}{1-\alpha} \quad (5)$$

Rényi entropy is a monotonically increasing function of Tsallis entropy.

Conditional entropies and statistical independence

Conditional entropies are just the ordinary entropies of the conditional probability distributions. Thus they can be calculated by establishing analogies with marginal distributions.

If the conditional probability function of X given $Y = y_j$ is represented by $f_{x/y_j}(x/y_j)$, then one can calculate conditional entropies by using Shannon's definitions, easily:

$$H(X/Y = y_j) = \sum_{i=1}^{n} f_{x_i/y_j}(x_i/y_j)\log(f_{x_i/y_j}(x_i/y_j)) \quad (6)$$

and the average conditional entropy can be calculated by

$$H(X/Y) = \sum_{j=1}^{m} f_{y_j}(y_j) H(X/Y = y_j) \quad (7)$$

For other conditional entropies (3) and (4) can be modified by using conditional probabilities instead of marginal probabilities. So by reasoning this way, Rényi and Tsallis conditional entropies can be calculated by the following equations:

$$H_R^{\alpha}(X/Y) = \sum_{j=1}^{m} f_Y(y_j)\left(\frac{1-log\sum_{i=1}^{n} f_{x_i/y_j}^{\alpha}(x_i/y_j)}{1-\alpha}\right) \quad (8)$$

$$H_T^{\alpha}(X/Y) = \sum_{j=1}^{m} f_Y(y_j)\left(\frac{1-\sum_{i=1}^{n} f_{x_i/y_j}^{\alpha}(x_i/y_j)}{\alpha-1}\right) \quad (9)$$

In this study, we will use (8) and (9) for measuring associations and/or correlations between variables. It should also be noted that there is still an ongoing controversy on defining conditional Tsallis and Rényi entropies. One can refer to Manije et al. (2013) for this purpose.

It is important to note that when the variables X and Y are independent, the average of conditional entropies are equal to marginal entropies. For two independent random variables;

$H(X/Y) = H(X)$ (10)

$H(Y/X) = H(Y)$ (11)

both of which mean that the occurrence of one of the two independent events $\{X = x_r\}$ and $\{Y = y_k\}$ does not give any information about the entropy of the probability distribution of the other.

A measure of mutual information

Several measures of correlation are used to investigate the dependencies between variables especially when the observations are from metric (i.e., interval or ratio) scale. When the variables belong to nominal or ordinal families, some alternative measures of association are used in literature. In addition to these, numerous measures based on entropy are also formulated, recently. Among these, mutual information that is based on relative entropy is used frequently.

Before discussing mutual information, the objective meaning of information should be noted: “In order to avoid possible misunderstandings, it should be emphasized that when we speak about information, what we have in mind is not the subjective “information” possessed by a particular observer. It is an uncertainty due to the fact that really several possibilities are to be taken into account. The measure of uncertainty does

not depend on anything else than these possible events and in the sense that it is entirely objective" (Rényi, 2007a).

The average amount of information conveyed by the random variable Y about X as once was suggested by C. E. Shannon is defined as

$$I(X,Y)=\sum_{j=1}^{m}\sum_{i=1}^{n} P\left(X=x_i, Y=y_j\right)\ log\frac{P\left(X=x_i, Y=y_j\right)}{P\left(X=x_i\right)\left(Y=y_j\right)} \quad (12)$$

Here the random variables are both discrete. Analogies to continuous cases are also straightforward. If X and Y are independent, mutual information is zero which means that the average amount of information given by one random variable about the other should be zero. It should be noted that $I(X,Y)\geq 0$ while the mutual information quantities may become negative for some pairs of events $\{X=x_i\}$ and $\{Y=y_j\}$ in sample space S. By direct application of the defining equations it can be shown that

$$I(X,Y)=H(X)+H(Y)-H(X,Y) \quad (13)$$

$$I(X,Y)=H(X)-H(X/Y) \quad (14)$$

$$I(X,Y)=H(Y)-H(Y/X) \quad (15)$$

Applications of mutual information

In many applications, one wants to maximize mutual information which is often equivalent to minimizing average conditional entropy. In telecommunications, the channel capacity is equal to the mutual information. Besides, it has also been used as a criterion for feature selection in machine learning. It is also used in determining the similarity of two different clustering of a data set as well as in Bayesian networks. For the context and statistical applications of entropy, relative entropy and mutual information, one can refer to Kullback(1997).

Variations of mutual information

The absolute value of Pearson (or linear) correlation coefficient is between zero and one. Here, zero means lack of linear correlation and absolute one means that there exists perfect linear relation between the variables.

In literature, there are other measures proposed like correlation ratio to detect nonlinear relationships between variables. Yet the problem with correlation ratio arises

from the difficulty for finding an asymptotic sampling distribution which is also true for linear correlation coefficient.

Suppose (X,Y) has a joint probability distribution. The two correlation ratios (eta squares) are defined as

$$\eta^2_{Y/X} = 1 - \frac{E(Var(Y/X))}{Var(Y)} \quad (16)$$

$$\eta^2_{X/Y} = 1 - \frac{E(Var(X/Y))}{Var(X)} \quad (17)$$

These two are not necessarily identical since X and Y cannot be used interchangeably.

It can further be said that

$0 \leq r^2 \leq \eta^2 \leq 1$ where r is the linear correlation coefficient.

The difference between η^2 and r^2 is used to detect nonlinear relationship between two variables X and Y, if it exists.

It should be also noted that Pearson correlation coefficient and eta-square measure the dependencies between variables belonging to metric scale. For nominal and ordinal scale, other association measures based on cell frequencies (e.g. mutual information) are to be used.

The problem with mutual information is that it does not necessarily have a definite upper bound though the lower bound is zero. For this reason some modifications on mutual information are realized by "normalizing". "Normalizing" is indeed, multiplying the statistic by some proper constant to ensure that the transformed version of the original statistic is now between zero (implying that the lack of one characteristic) and one (implying the existence of one characteristic).

Normalized variants of the mutual information are provided by the coefficients of constraint (Coombs et al., 1970) or uncertainty coefficient (Press & Flannery; 1988)

$$C_{XY} = \frac{I(X,Y)}{H(Y)} \quad (18)$$

$$C_{YX} = \frac{I(X,Y)}{H(X)} \quad (19)$$

The two coefficients are not necessarily equal. A more useful and symmetric one is

$$R(X,Y) = \frac{I(X,Y)}{H(X)+H(Y)} \quad (20)$$

This is redundancy and gets a value of zero when the variables are independent and a maximum value when one variable becomes completely redundant with the knowledge

of the other. To be more specific it is going to be ½, if the two variables are correlated perfectly. This result implies that one of the two perfectly correlated variables is unnecessary or redundant.

In literature there are still some other measures for dependencies:

$$\frac{I(X,Y)}{min\{H(X),H(Y)\}} \quad (21) \quad , \quad \frac{I(X,Y)}{H(X,Y)} \quad (22) \quad , \quad \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \quad (23)$$

(Yao (2003), Strehl & Ghosh (2002)

For other association measures based on other entropy measures, in this paper we have started with basic definitions given in (14) and (15), but rather used Rényi and Tsallis formulations given in (3) and (4) . Then to normalize, we used the following formula:

$$U(X,Y) = \frac{H(Y)-H(Y/X)}{H(Y)} \quad (24)$$

which is the analogous form of eta-square in (16) It is also equal to C_{XY} in (17). It is also possible to derive the other uncertainty coefficient as

$$U(Y,X) = \frac{H(X)-H(X/Y)}{H(X)} \quad (25)$$

which is equal to C_{YX} in (19)

3. Application on Metric Data

Just for illustrative purposes, we considered three different sets of hypothetical metric data to make meaningful comparisons between information statistics and classic linear correlation coefficients and correlation ratios.

A quick inspection on the table below will reflect high positive correlation between X and Y variables, since the largest probabilities appear on the main diagonal of the contingency table.

Table 1. A fictitious probability distribution of two highly and positively correlated metric variables X and Y

Distribution 1

X/Y	0	1	2
0	0.19	0.005	0.005
1	0.005	0.39	0.005
2	0.005	0.005	0.39

Note that the square root of $\eta^2_{Y/X}$ is denoted by "Eta1(Y/X) and that of $\eta^2_{X/Y}$ is denoted by "Eta2(X/Y) in the following table.

Table 2. Classic correlation measures and information measures based on Shannon entropy

Cor(X,Y)	Eta1(Y/X)	Eta2(X/Y)	Shannon	H(X,Y)	H(X)	H(Y)	I(X,Y)	U(X,Y)	U(Y,X)
0.94643	0.9465	0.9465	1.52193	1.7441	1.522	1.522	1.2997	0.854	0.854

A rapid check on Table 2. may suggest that classic correlation measures and other association measures based on entropy are in harmony to a certain degree.

Different parametrizations on Rényi and Tsallis entropies yield the following table suggesting similar conclusions.

Table 3. Rényi and Tsallis mutual information statistics under different parametrizations

Alpha	Entropy	H(X,Y)	H(X)	H(Y)	H(X/Y)	H(Y/X)	I(X,Y)	U(X,Y)	U(Y,X)
0.99	Rényi	1.74849	1.52248	1.5225	0.068	0.068	1.2965	0.852	0.8516
0.99	Tsallis	1.21933	1.06089	1.0609	0.157	0.157	0.9025	0.851	0.8507
1.5	Rényi	1.61157	1.4961	1.4961	0.036	0.036	1.3806	0.923	0.9228
1.5	Tsallis	0.8559	0.80919	0.8092	0.082	0.082	0.7625	0.942	0.9423
2	Rényi	1.55449	1.47393	1.739	0.026	0.026	1.3934	0.945	0.9453
2	Tsallis	0.65955	0.64	0.64	0.059	0.059	0.6205	0.969	0.9695
2.5	Rényi	1.52084	1.45508	1.4551	0.022	0.022	1.433	0.955	0.9548
2.5	Tsallis	0.52952	0.51982	0.5198	0.049	0.049	0.4711	0.981	0.9813
3	Rényi	1.49713	1.43916	1.4392	0.02	0.02	1.3812	0.96	0.9597
3	Tsallis	0.43725	0.432	0.432	0.044	0.044	0.4267	0.988	0.9878

As a second situation, we considered the following joint distribution in which the variables are nearly uncorrelated.

Table 4. A fictitious probability distribution of probably uncorrelated two metric variables X and Y

Distribution2

X/Y	0	1	2
0	0.2	0.05	0.15
1	0.05	0.1	0.05
2	0.05	0.3	0.05

The summarizing association and correlation measures for this bivariate scheme is as follows:

Table 5. Classic correlation measures and information measures based on Shannon entropy for X and Y variables that seem possibly uncorrelated.

Cor(X,Y)	Eta1(Y/X)	Eta2(X/Y)	Shannon	H(X,Y)	H(X)	H(Y)	I(X,Y)	U(X,Y)	U(Y,X)
0.07555	0.08276	0.56335	0.52193	2.8087	1.522	1.539	0.2527	0.166	0.1642

It should also be emphasized that for all tables given below, all mutual information measures are calculated by the help of the equations (14) and (15). Since there is not a widely accepted definition of relative entropy, one point gains importance: By adopting basic definitions, mutual information can be evaluated as the difference between the entropy of a marginal distribution and the average conditional entropy.

Table 6. Rényi and Tsallis mutual information statistics under different parametrizations for uncorrelated X and Y variables.

Alpha	Entropy	H(X,Y)	H(X)	H(Y)	H(X/Y)	H(Y/X)	I(X,Y)	U(X,Y)	U(Y,X)
0.99	Rényi	2.81214	1.52248	1.5399	0.383	0.388	0.2503	0.164	0.1625
0.99	Tsallis	1.96835	1.06089	1.0731	0.886	0.898	0.1657	0.156	0.1544
1.5	Rényi	2.64834	1.4961	1.5166	0.343	0.355	0.3644	0.244	0.2403
1.5	Tsallis	1.20124	0.80919	0.8176	0.652	0.666	0.4256	0.526	0.5205
2	Rényi	2.51457	1.47393	1.4941	0.313	0.332	0.4535	0.308	0.3035
2	Tsallis	0.825	0.64	0.645	0.512	0.525	0.46	0.719	0.7132
2.5	Rényi	2.40719	1.45508	1.4722	0.289	0.316	0.5201	0.357	0.3533
2.5	Tsallis	0.6121	0.51982	0.5224	0.42	0.43	0.4301	0.827	0.8234
3	Rényi	2.32193	1.43916	1.4512	0.271	0.304	0.5684	0.395	0.3917
3	Tsallis	0.48	0.432	0.4331	0.356	0.363	0.3851	0.891	0.8892

Unfortunately for this situation, we have to be very careful when analysing dependencies between variables by using these two parametric representations. Because entropy estimates by either Rényi and Tsallis methods inevitably depend on the chosen alpha values. Especially, in this example we see that Tsallis entropies are very sensitive to different parametrizations compared to Rényi and Shannon entropies. So different parametrizations may yield misleading results.

As the final situation we consider the following table:

Table 7. A fictitious probability distribution of two highly but negatively correlated metric variables X and Y

Distribution3

X/Y	0	1	2
0	0.01	0.01	0.38
1	0.01	0.28	0.01
2	0.28	0.01	0.01

Classic correlation and association measures are given below:

Table 8. Classic correlation measures and information measures based on Shannon entropy for X and Y variables that seem possibly highly but negatively correlated.

Cor(X,Y)	Eta1(Y/X)	Eta2(X/Y)	Shannon	H(X,Y)	H(X)	H(Y)	I(X,Y)	U(X,Y)	U(Y,X)
-0.913	0.91307	0.91307	1.57095	1.9575	1.571	1.571	1.1844	0.754	0.7539

The summarizing statistics for other association measures are as follows:

Table 9. Rényi and Tsallis mutual information statistics under different parametrizations for uncorrelated X and Y variables.

Alpha	Entropy	H(X,Y)	H(X)	H(Y)	H(X/Y)	H(Y/X)	I(X,Y)	U(X,Y)	U(Y,X)
0.99	Rényi	1.96258	1.57109	1.5711	0.118	0.118	1.1796	0.751	0.7508
0.99	Tsallis	1.36966	1.09495	1.095	0.272	0.272	0.8202	0.749	0.7491
1.5	Rényi	1.79631	1.56372	1.5637	0.071	0.071	1.3311	0.851	0.8513
1.5	Tsallis	0.92686	0.83677	0.8368	0.156	0.156	0.7467	0.892	0.8923
2	Rényi	1.72834	1.55639	1.5564	0.053	0.053	1.3845	0.89	0.8895
2	Tsallis	0.6982	0.66	0.66	0.115	0.115	0.6218	0.942	0.9421
2.5	Rényi	1.69276	1.549	1.549	0.045	0.045	1.4052	0.907	0.9072
2.5	Tsallis	0.55197	0.53348	0.5335	0.095	0.095	0.515	0.965	0.9653
3	Rényi	1.6698	1.54157	1.5416	0.04	0.04	1.4133	0.917	0.9168
3	Tsallis	0.45061	0.441	0.441	0.085	0.085	0.4314	0.978	0.9782

Here again, we can observe the parallelism between classic correlation measures and mutual information measures even under different parametrizations.

4. Results

In this study we have applied some relative entropy measures on some hypothetical data for detecting dependencies between two metric variables to check whether these measures agree with classic correlation measures in general. We have considered three hypothetical data (two bivariate distributions for highly correlated variables and one bivariate distribution for nearly uncorrelated variables).

Whenever the variables are highly correlated (whether positively or negatively), to some extent nearly all measures have produced similar results. The reason for this phenomenon may be explained intuitively as follows: Since the variables are highly correlated, the information about one variable decreases uncertainty of the other variable totally. In other words the difference between the entropy of a marginal distribution and the average conditional entropy is approximately equal to the entropy of the marginal distribution. The average conditional entropy will be nearly zero because of strong linear dependence. In this situation, mutual information is only a function of a marginal entropy.

The situation is different for uncorrelated variables because in this case mutual information is a function of two entropies. The parametrization effects on two entropies may or may not balance each other. Therefore uncorrelatedness may be a factor that increases uncertainty. In case of higher uncertainty, different mutual information measures behave differently.

Finally, Rényi entropies seem more stable under different parametrizations because the operations in definition (i.e., taking the powers of probabilities and then taking the logarithm of the sum of powers of probabilities) seem acting on opposite directions. Therefore association measures based on different parametrizations of Tsallis entropies may show more variability compared to Rényi entropies.

REFERENCES

Cahill, N. (2010). Normalized Measures of Mutual Information with General Definitions of Entropy for Multimodal Image Restoration. B.Fischer, B.Dawant, and C.Lorentz (Eds), WBIR 2010, LNCS 6204, 258-268, 2010, © Springer-Verlag, Berlin Heidelberg.

Coombs, C.H.,Daves,R.M.&Tversky(1970). Mathematical Psychology: An Elementary Introduction. Prentice-Hall,Englewood Cliffs, New Jersey.

Cover, T.M.; Thomas, J.A.(2006). Elements of Information Theory. Wiley Interscience (Second Edition), Hoboken, New Jersey.

Esteban, M.D.,Morales, D.(1995). A summary on Entropy Statistics. Kybernetika, Vol. 31(1995), 4, 337-346.

Golan, A.(2006), Information and Entropy Econometrics: A Review and Synthesis,Foundations and Trends in Econometrics. 2:1-2, Now Publishers Inc.

Kullback, S. (1997). Information Theory and Statistics. Dover Publications, New York.

Manije, S.T., Gholamreza, M.B., Mohammed, A. (2013). Conditional Tsallis Entropy. Cybernetics and Information Technologies, Volume 13, No.12, Bulgarian Academy of Sciences, 37-42.

Pardo, L (2006), Statistical Inference Measures Based on Divergence Measures. Chapman&Hall/CRC.

Press, W.H., Flannery, B.P., Teukolsky, S.A.,& Vetterling, W.T.(1988). Numerical Recipes in C :The Art of Scientific Computing. Cambridge University Press, Cambridge, 634.

Rényi, A.(2007a). Probability Theory. Dover Publications, New York, 554.

Rényi, A.(2007b). Foundations of Probability. Dover Publications, New York.

Reza,F.M.(1994). An Introduction to Information Theory. Dover Publications, New York.

Strehl,A.,Ghosh(2002). Cluster ensembles—a Knowledge Reuse Framework for Combining Multiple Participations. (http://strehl.com/download/strehl-jmlr02.pdf) (PDF) , Journal of Machine Learning Research 3: 583-617.

Yao, Y.Y.(2003). Information-theoretic Measures for Knowledge Discovery and Data Mining in Entropy Measures. Maximum Entropy Principle and Emerging Applications, Karmeshu (ed.) Springer, 115-136.

Ullah, A. (1996). Entropy, Divergence and Distance Measures with Econometric Applications. Journal of Statistical Planning and Inference 49 (1996), Elsevier.

ELECTRONIC SOURCES

http://en.wikipedia.org/wiki/Ludwig_Boltzmann#cite_note-11
http://en.wikipedia.org/wiki/Mutual_information
http://scienceworld.wolfram.com/physics/BoltzmannEquation.html