Nonparametric Estimation of Mutual Information and Test for Independence

Deemat C Mathew

University of Hyderabad, Andhra Pradesh, INDIA.

Corresponding Address:

deematcm@yahoo.com

Research Article

Abstract: Mutual information is a concept in information theory which may help us to define independence. We present estimators of mutual information of continuous random variables with heavy tails based on histogram and describe their asymptotic properties. Under appropriate assumptions on the tail behavior of the random variables, we obtain root N consistency of the estimators. We analyses the usefulness of this measure in testing statistical dependence.

Keywords: Mutual information, Entropy, Kernel density, Independence.

1. Introduction

Measuring dependence between the random variables is a fundamental and interesting problem in statistics. It finds applications in all the fields of statistics and is very useful in time series analysis. There is lot of attention in obtaining a measure that describes the dependence between the random variables. The classical and most popular measure of linear dependence is the correlation coefficient. It is commonly used in many areas due to its simplicity, low computational cost and ease of estimation. But, it is well known that correlation is not equivalent to dependence. Correlation cannot be defined for models with stable or heavy tail distributions due to the lack of second order moments (Adler et al. (1998), Brockwell and Davis (1987)). Granger (1983) illustrated that some bilinear and deterministic chaotic series have all correlation zero. Mutual information between two random variables which describes the reduction in uncertainty of one random variable knowing the other random variable. Mutual information is same as the Kullback-Leibler divergence between the joint probability density and product of marginal probability densities. Mutual information has the property that it is always non negative and is zero if and only if the random variables are independent. We can extend this to define conditional mutual information between two random variables given another. This property provides possibility of using mutual information and as dependence measures. Another measure of dependence is the *informational coefficient of correlation* introduced by Linfoot (1957), which is defined for continuous random variables.This measure is an increasing monotone function of mutual information and it preserves the attractive properties of mutual information. Also it lies between zero and one.This property is a useful standardization when comparing different dependence measures. After Linfoot's initial work on this measure, more of its properties were studied and applied by Granger and Lin (1994) and Dionisio et al. (2004). In the present study we consider estimation of mutual information of continuous random variables with heavy tails based on histogram. Since mutual information is functions of entropy, the estimation demands the estimation of functionals of probability density functions (Moddemeijer (1989)). Kernel density estimates are widely used nonparametric technique to estimate probability density function (Wand and Jones (1995)). We use histogram kernel to estimate probability density functions as its asymptotic properties are well studied (Ahmad and Lin (1976), Joe (1989)). We obtain asymptotic properties under some assumptions on the tail of the distribution. The rest of the paper is organized as follows, In Section 2; we introduce entropy and mutual information measures and study its properties. Estimation of mutual information is discussed in Section 3. Histogram based estimation and its asymptotic properties are studied in Section 4. Testing independence and a bootstrap algorithm for testing are discussed in the last section.

2. Entropy and Mutual Information

In this section we start with a brief overview of information theoretic measures required for subsequent development of the theory. After introducing entropy and mutual information we discuss the properties that help us to measure dependence.

Definition 1:

Let X and Y be absolutely continuous random variable with joint density function $f_{X,Y}(x,y)$ and marginals $f_X(x)$ and $f_Y(\mathrm{y})$ respectively. The information contained in X given by the Shannon entropy is defined as

$$H(X) = -\int f_X(x) \log f_X(x)\, dx. \qquad (1)$$

The joint entropy of (X,Y) is given by

$$H(X,Y) = -\int f_{X,Y}(x,y) \log f_{X,Y}(x,y)\, dxdy. \qquad (2)$$

The conditional entropy of X given Y is defined as

$$H(X \mid Y) = -\int\int f_{X,Y}(x,y) \log f(x \mid y), \qquad (3)$$

where $f(x \mid y)$ is the conditional distribution of x given y.

If X and Y are independent, then H(X,Y)=H(X)+H(Y) and H(X|Y)=H(X).

Definition 2:

The mutual information I(X,Y) is defined as

$$I(X,Y) = \int\int f(x,y) \log \frac{f(x,y)}{f(x) f(y)} dxdy. \qquad (4)$$

Simple algebra shows that

$$\begin{aligned} I(X,Y) &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Remark 1:

I(X,Y) is a symmetric measure of dependence between X and Y. It is also the Kulback-Leibler distance between joint density $f_{X,Y}(x,y)$ and the product of marginal densities $f_X(x)$ and $f_Y(\mathrm{y})$. Hence I(X,Y) is always non negative and zero if and only if X and Y are independent.

Remark 2:

$I(X,Y) = +\infty$. This follows from data processing theorem (Cover and Thomas (1991)). Note that if X a discrete random variable, then I(X,X)= H(X), the entropy of X.

Next we study some properties of these measures.

Theorem 1

I(X,Y)$ is invariant to separate one to one transformation.

Proof:

Let $X^* = h_1(X)\, and\, Y^* = h_2(Y)$.

Let g_{12} be the joint density of X^* , Y^*, g_1 and g_2 be the densities of X^* and Y^* respectively.

Consider

$$\begin{aligned} I(X^*,Y^*) &= \int \log\Big(\frac{g_{12}(x^*,y^*)}{g_1(x^*) g_2(y^*)}\Big) g_{12}(x^*,y^*) dx^* dy^* \\ &= \int \log\Big(\frac{f_{X,Y}(h_1^{-1}x^*, h_2^{-1}y^*)}{f_X(h_1^{-1}x^*) f_Y(h_2^{-1}y^*)}\Big) \quad f_{X,Y}(h_1^{-1}x^*, h_2^{-1}y^*) \frac{dh_1^{-1}}{dx^*} \frac{dh_2^{-1}}{dy^*} dx^* dy^*. \end{aligned}$$

Or

$$\begin{aligned} I(X^*,Y^*) &= \int \log\Big(\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y}\Big) f_{X,Y}(x,y) dxdy \\ &= I(X,Y). \end{aligned}$$

This completes the proof.

Now we will discuss the estimation of mutual information.

3. Estimation of Mutual information

The estimation of mutual information is equivalent to estimation of the functional of density. The natural estimate are

$$\overline{I}(X,Y)=\frac{1}{n}\sum_{t=1}^{n}\log\frac{f(x,y)}{f(x)f(y)}, \qquad (6)$$

which is consistent for I(X,Y) when $\int f(x,y)(\log\frac{f(x,y)}{f(x)f(y)}dxdy<\infty$.

Unfortunately it is not practical to use the above estimate when the underlying density f itself is unknown. In this scenario the estimation procedure involves the following steps: First estimate the joint density/conditional density. After obtaining the density estimate, evaluate the integral to obtain the estimate $\hat{I}(X,Y)$.

For a given kernel function K(.) we can estimate the p-variate density nonparametrically as

$$\hat{f}_i(x)=\frac{1}{(n-1)h^p}\sum_{j=1,\,j\neq i}^{n}K\{(x-X_j)/h\} \qquad (7)$$

or

$$\hat{f}(x)=\frac{1}{nh^p}\sum_{j=1}^{n}K\{(x-X_j)/h\}, \qquad (8)$$

and we can estimate

$$\hat{I}(X,Y)=\frac{1}{n}\sum_{t\in S}log\frac{\hat{f}(x_t,y_t)}{\hat{f}(x_t)\hat{f}(y_t)}. \qquad (9)$$

The set S is a subset of (1,2,...,n) and is introduced in case is necessary to trim the summands.

Now we establish the consistency of the estimator of $I(X,Y)$.

Let $(X_1,Y_1),(X_2,Y_2),...(X_n,Y_n)$ be a sample from bivariate distribution $f(x,y)$ with marginals' f(x) and f(y).

Suppose $E|\log f(X,Y)|<\infty$ and each $E|\log f(X)|<\infty$ and $E|\log f(Y)|<\infty$.

Suppose $\hat{f}(x,y)$, $\hat{f}(x)$ and $\hat{f}(y)$ are density estimators of $f(x,y)$, $f(x)$ and $f(y)$ such that

$$\max_{t\in S}\left|\frac{\hat{f}(x_t,y_t)}{f(x_t,y_t)}-1\right|\xrightarrow{P}0,$$

$$\max_{t\in S}\left|\frac{\hat{f}(x_t)}{f(x_t)}-1\right|\xrightarrow{P}0, \qquad (A)$$

and $\max_{t\in S}\left|\frac{\hat{f}(y_t)}{f(y_t)}-1\right|\xrightarrow{P}0$,

Theorem

If the above assumptions (A) holds then, $\hat{I}(X,Y)\xrightarrow{P}I(X,Y)$.

Proof

Consider

$$\begin{aligned}\hat{I}(X,Y)&=\frac{1}{n}\sum_{t\in S}\log\frac{\hat{f}(x_t,y_t)}{\hat{f}(x_t)\hat{f}(y_t)}\\&=\frac{1}{n}\sum_{t\in S}\log\frac{\hat{f}(x_t,y_t)f(x_t)f(y_t)}{\hat{f}(x_t)\hat{f}(y_t)f(x_t,y_t)}+\frac{1}{n}\sum_{t\in S}\log\frac{f(x_t,y_t)}{f(x_t)f(y_t)}\\&=\frac{1}{n}\sum_{t\in S}\log\frac{\hat{f}(x_t,y_t)}{f(x_t,y_t)}-\frac{1}{n}\sum_{t\in S}\log\frac{\hat{f}(x_t)}{f(x_t)}-\frac{1}{n}\sum_{t\in S}\log\frac{\hat{f}(y_t)}{f(y_t)}+\frac{1}{n}\sum_{t\in S}\log\frac{f(x_t,y_t)}{f(x_t)f(y_t)}\end{aligned}$$

Now since $|\log(1+x)|\leq 2|x|$ by assumptions (A) the first three terms converges to 0 in probability and by ergodic theorem

$$\frac{1}{n}\sum_{t\in S}\log\frac{f(x_t,y_t)}{f(x_t)f(y_t)}\xrightarrow{P}E(\log\frac{f(X,Y)}{f(X)f(Y)})$$
$$=I(X,Y).$$

Now in the next section we will show how the estimator of mutual information helps to test independence.

4. Test for Independence

From Remark 2, two random variables X and Y are independent if and only if $I(X,Y)=0.$ In practise as we are using sample estimate of mutual information, we want to test the significance of $\hat{I}(X,Y).$ In the present study we adopt a bootstrap approach to test the significance of $\hat{I}(X,Y).$ Under the null hypothesis we assume that the random variables are independent. We estimate the achieved significance level (ASL) through bootstrap samples of the original time series. The test procedure is thus composed of the following steps:

1. Calculate $\hat{I}(X,Y)$ for the observed sample $\left(x_t,y_t\right)\quad t=1,2,...,n.$
2. Randomly permute the sample $\{(x_t,y_t)\}$ and obtain a bootstrap sample $(\tilde{x}_t,\tilde{y}_t)$ of same size n.
3. Calculate $\tilde{I}(X,Y)$ for the bootstrap sample $(\tilde{x}_t,\tilde{y}_t)$.
4. Repeat steps 2-3 B times.
5. Calculate the one-sided bootstrap ASL as

$$\hat{p}(k)=\frac{1+\sum_{j=1}^{B}I(\tilde{I}(X,Y)\geq\hat{I}(X,Y))}{1+B}.$$

6. Reject the null hypothesis of independence if $\hat{p}(k)\leq\alpha,$ where α denotes the chosen significance level.

5. Conclusion

In the present study we discussed the estimation of mutual information and studied its asymptotic properties. We showed that the estimator can be used for testing statistical dependence between pair of random variables and bootstrap procedure for testing is also given.

References

1. Adler, R., Feldman, R., and Taqqu, M. (1998). A practical guide to heavy tails, Birkh¨auser.
2. Ahmad, I. A. and Lin, P. E., (1976). A nonparametric estimation of the entropy for absolutely continuous distributions, IEEE Trans. Inf. Theory, 22, 372-375.
3. Brillinger, D.R., (2004). Some data analyses using mutual information. Brazilian J. Probab. Statist. 18, 163–183.
4. Brockwell, P.J. and Davis, R.A. (1991). Time Series: Theory and Methods, 2nd ed., Springer, NewYork.
5. Cover, T. M. and Thomas J. A., (1991) Elements of Information Theory. Wiley.
6. Darbellay, G., (1998). An adaptive histogram estimator for mutual information. UTIA Research Report 1889. Academy of Sciences, Prague.
7. Granger, C. W. J., (1983). Forecasting whitenoise. In Zellner, A. (ed.), Applied Time Series Analysis of Economic Data, pp. 308–314. Washington, DC: Bureau of the Census.
8. Granger, C. and J.-L. Lin.,(1994). Using the Mutual Information Coefficient to Identify Lags in Nonlinear Models, J. Time Ser. Anal., 15(4), 371-384.
9. Joe, H. (1989). Estimation of entropy and other functionals of multivariate density, Ann. Inst.Statist. Math., 4, 683-697.
10. Hall, P. and Morton S. C., (1993). On the estimation of the entropy, Ann. Inst. Statist. Math., 45, 69-88.
11. Moddemeijer, R., (1989). On estimation of entropy and mutual information of continuous distributions, Signal Process, 16, 233-248.
12. Renya, A., (1959). On measure of dependence, Acta Mathematica Academiae Scientiarium Hungaricae,10, 441-451.
13. Silverman, B., (1986).Density Estimation for Statistics and Data Analysis,\ Chapman and Hall. London.