Journal of Mathematical Research & Exposition
Jan., 2010, Vol. 30, No. 1, pp. 87–98
DOI:10.3770/j.issn:1000-341X.2010.01.008
Http://jmre.dlut.edu.cn

Detecting Lags in Nonlinear Models Using General Mutual Information

Wei GAO[1,2,*], **Zheng TIAN**[1,3]

1. *Department of Applied Mathematics, Northwest Polytechnical University, Shaanxi* 710072*, P. R. China;*
2. *School of Statistics, Xi'an University of Finance & Economics, Shaanxi* 710061*, P. R. China;*
3. *National Key Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080*, P. R. China*

Abstract The general mutual information (GMI) and general conditional mutual information (GCMI) are considered to measure lag dependences in nonlinear time series. Both of the measures have the property of invariance with transform. The statistics based on GMI and GCMI are estimated using the correlation integral. Under the hypothesis of independent series, the estimators have Gaussian asymptotic distributions. Simulations applied to generated nonlinear series demonstrate that the methods appear to find frequently the correct lags.

Keywords general mutual information; general conditional mutual information; nonlinear time series; lag dependence.

Document code A
MR(2000) Subject Classification 62M10
Chinese Library Classification O211.6

1. Introduction

Statistics based on mutual information have been found to be helpful in determining the lags in nonlinear time series, using the identification methods discussed by Granger and Lin [1, 2] and others [3, 4]. However, the appropriate statistics to distinguish direct and indirect dependence are conditional mutual information measures. These statistics based on Shannon entropy involve high dimension kernel density estimation, which is computationally costly to achieve reasonable accuracy. Diks and Manzan [5] used the correlation integrals to estimate conditional mutual information. In this paper we consider statistics based on Renyi entropy, which takes Shannon entropy as a special case. We also prove that the statistics satisfy the properties for being statistics to measure bivariate dependence. The statistics based on general mutual information are estimated by correlation integral, which are faster than estimates of the probability density function. Under the hypothesis of independent series, the estimators have Gaussian asymptotic distributions, while there is no similar result for kernel density estimate. The methods are applied

Received December 11, 2007; Accepted January 5, 2009
Supported by the National Natural Science Foundation of China (Grant Nos. 60375003; 60972150) and the Science and Technology Innovation Foundation of Northwestern Polytechnical University (Grant No. 2007KJ01033).

[*] Corresponding author

E-mail address: gaoww525@tom.com (W. GAO)

to simulated nonlinear time series. The results show that the statistic based on GMI behaves much like that based on MI, and the statistic based on GCMI behaves better than the statistic in [1] for identifying correct lags.

The rest of the paper is organized as follows. The measures based on information theory are presented and the properties of the statistics are proved in Section 2. Section 3 gives the nonparametric estimation based on correlation integrals and the asymptotic distribution of the estimator. In Section 4, numerical examples are presented. Finally conclusions are summarized in Section 5.

2. The general mutual information

Suppose that a pair of continuous random variables X, Y has a joint probability density function $f_{X,Y}(x,y)$ with marginals $f_X(x)$, $f_Y(y)$, respectively. The information contained in X is given by the Shannon entropy

$$H(X) = -\int [\ln f_X(x)] f_X(x)\mathrm{d}x. \tag{1}$$

The amount of information jointly contained in the random vector (X, Y) is measured by the joint entropy

$$H(X,Y) = -\iint [\ln f_{X,Y}(x,y)] f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y. \tag{2}$$

To quantify the dependence between variables, a commonly used measure is the mutual information

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{3}$$

The mutual information is non-negative, and $I(X;Y) = 0$ if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

When modelling for a time series X_t, we refer to establish a regressive model of explanatory variable X_t on explained variables $X_{t-1}, \ldots, X_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$. In the case of time series, X_t is known as current variable, while $X_{t-1}, \ldots, X_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are known as lagged variables. Then the problem of selecting associate explained variables can also be called lag selection.

Granger [1] used $R(X,Y) = [1 - \exp\{-2I(X;Y)\}]^{1/2}$ to identify what lags to use when building nonlinear models. The estimation of R is based on the kernel density estimate of the probability density function, which is computationally costly.

Instead of using the Shannon entropy, a generalization of the statistic is obtained by using the Renyi entropies

$$H_q(X) = -\frac{1}{q-1} \ln \int (f_X(x))^{q-1} f_X(x)\mathrm{d}x. \tag{4}$$

It is easy to show that the limit as $q \longrightarrow 1$ leads to the Shannon entropy. The q-order mutual information is defined by

$$I_q(X;Y) = H_q(X) + H_q(Y) - H_q(X,Y). \tag{5}$$

The q-order conditional mutual information, quantifying the average amount of additional information in Y about X, given the information about X already contained in Z, is defined by

$$I_q(X;Y,Z) - I_q(X;Z) = -H_q(X,Y,Z) + H_q(X,Z) + H_q(Y,Z) - H_q(Z) = I_q(X;Y|Z). \quad (6)$$

In this paper, we mainly consider the 2-order Renyi entropy. From Equ.(4), the condition for $H_2(X)$ to be well-defined and finite is that the probability density function $f_X(x)$ is square integrable on the probability space. Throughout the paper, we assume that the process X satisfies the integrable condition.

The properties of statistic $I_q(X;Y|Z)$ are shown in the following theorems. With similar method, one can prove that the properties also exist for $I_q(X;Y)$.

Theorem 1 *$I(X;Y|Z)$ is nonnegative and is 0 if and only if X and Y are statistically independent given Z.*

Proof Kullback-Leibler information divergence

$$I(f_1;f_2) = \int \ln \frac{f_1(x)}{f_2(x)} f_1(x)\mathrm{d}x$$

is always nonnegative and $I(f_1;f_2) = 0$ if and only if $f_1 = f_2$ (see [6]).

The conditional mutual information between X, Y for given Z is defined as

$$I(X;Y|Z) = \iiint f_{X,Y,Z}(x,y,z) \ln \frac{f_{X,Y,Z}(x,y,z)}{f_{X|Z}(x|z)f_{Y|Z}(y|z)f_Z(z)} \mathrm{d}x\mathrm{d}y\mathrm{d}z,$$

where $f_{X|Z}(x|z)$ is the conditional density function of X given Z. The result follows immediately from the properties of Kullback-Leibler information divergence, where $f_1 = f_{X,Y,Z}(x,y,z)$, $f_2 = f_{X|Z}(x|z)f_{Y|Z}(y|z)f(z)$. $\square$

Theorem 2 $I_q(X;Y|Z)$ *is invariant to separate one-to-one transformations.*

Proof Without loss of generality, it will be assumed that the one-to-one transformations h_1, h_2, h_3 are differentiable. Let $X^* = h_1(X)$, $Y^* = h_2(Y)$, $Z^* = h_3(Z)$ and g, g_{13}, g_{23}, g_3 be the joint and marginal densities for X^*, Y^*, Z^*. By definition,

$$\begin{aligned}
I_q(X^*;Y^*|Z^*) =& \frac{1}{1-q} \ln \frac{\int g_{13}^q(X^*,Z^*)\mathrm{d}x^*\mathrm{d}z^* \int g_{23}^q(Y^*,Z^*)\mathrm{d}y^*\mathrm{d}z^*}{\int g^q(X^*,Y^*,Z^*)\mathrm{d}x^*\mathrm{d}y^*\mathrm{d}z^* \int g_3^q(Z^*)\mathrm{d}z^*} \\
=& \frac{1}{1-q} \ln \frac{\int f_{13}^q \left\{h_1^{-1}(X^*), h_3^{-1}(Z^*)\right\} \frac{dh_1^{-1}}{\mathrm{d}x^*} \frac{\mathrm{d}h_3^{-1}}{\mathrm{d}z^*} \mathrm{d}x^*\mathrm{d}z^*}{\int f^q \left\{h_1^{-1}(X^*), h_2^{-1}(Y^*), h_3^{-1}(Z^*)\right\} \frac{\mathrm{d}h_1^{-1}}{\mathrm{d}x^*} \frac{\mathrm{d}h_2^{-1}}{\mathrm{d}y^*} \frac{\mathrm{d}h_3^{-1}}{\mathrm{d}z^*} \mathrm{d}x^*\mathrm{d}y^*\mathrm{d}z^*} + \\
& \frac{1}{1-q} \ln \frac{\int f_{23}^q \left\{h_2^{-1}(Y^*), h_3^{-1}(Z^*)\right\} \frac{\mathrm{d}h_2^{-1}}{\mathrm{d}y^*} \frac{\mathrm{d}h_3^{-1}}{\mathrm{d}z^*} \mathrm{d}y^*\mathrm{d}z^*}{\int f_3^q \left\{h_1^{-1}(Z^*)\right\} \frac{dh_3^{-1}}{\mathrm{d}z^*} \mathrm{d}z^*} \\
=& \frac{1}{1-q} \ln \frac{\int f_{13}^q(X,Y,Z)\mathrm{d}x\mathrm{d}z \int f_{23}^q(Y,Z)\mathrm{d}y\mathrm{d}z}{\int f^q(X,Y,Z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \int f_3^q(Z)\mathrm{d}z} \\
=& I_q(X;Y|Z). \quad \square
\end{aligned}$$

Theorem 3 *Consider vector $X = (X_1, X_2, X_3)$ with Gaussian distribution $N_3(\mu, V)$. Let $\rho_{12|3} = \mathrm{corr}(X_1, X_2|X_3)$, i.e., the partial correlation coefficient between X_1 and X_2 given X_3.*

Then $I_q(X_1;X_2|X_3) = -\frac{1}{2}\ln(1-\rho^2_{12|3})$.

Proof In this special case of Gaussian vector, the joint entropy of Equ.(4) can be computed straightforwardly from the definition

$$\begin{aligned} H_q(X_1,X_2,X_3) &= -\frac{1}{q-1}\ln\Big(\frac{|V|^{q/2}}{(2\pi)^{3q/2}}\int e^{-\frac{q}{2}(x-\mu)'V^{-1}(x-\mu)}\mathrm{d}x\Big) \\ &= \frac{3}{2}\ln(2\pi)+\frac{1}{2}\ln|V|+\frac{3\ln q}{2q-1}. \end{aligned} \tag{7}$$

The q-order conditional mutual information of X_1 and X_2 given X_3 is

$$\begin{aligned} I_q(X_1,X_2|X_3) &= H_q(X_1,X_3)+H_q(X_2,X_3)-H_q(X_1,X_2,X_3)-H_q(X_3) \\ &= \frac{1}{2}\ln\Big(\frac{|V_{13}|\,|V_{23}|}{|V_3|\,|V|}\Big), \end{aligned} \tag{8}$$

where V_{13}, V_{23} are the covariance matrix of (X_1,X_3) and (X_2,X_3), respectively, V_3 is the variance of X_3.

Then the theorem is obvious from Proposition 6.4.6 in [6]. □

Remark 1 The theorem can be extended to the case that X_3 is a vector straightly.

Thus in Gaussian distributions the information measures for measuring the conditional independence of X_1 and X_2 given X_3 is a simple function of the partial correlation between X_1 and X_2 given X_3. Then the problem of test independence in linear model is a special situation of the information measures.

Strictly speaking, the general conditional mutual information $I_q(X;Y|Z)$ is not positive definite for $q \neq 1$. This means that it is possible to construct examples of variables X and Y, which are conditionally dependent given Z, and for which $I_q(X;Y|Z)$ is zero or negative. This situation appears to be very exceptional, and usually $I_2(X;Y|Z)$ is either positive or negative. This suggests that a one-sided test, rejecting for $I_2(X;Y|Z)$ large, is not always optimal. In practice, $I_2(X;Y|Z)$ behaves much like $I_1(X;Y|Z)$ in that we usually observe larger power for one-sided tests (rejecting for large $I_2(X;Y|Z)$) than for two-sided tests. This leads us to choose $q=2$, together with a one-sided implementation of the test. Pompe [7] proposed an approach to transform the data to have a uniform distribution to guarantee that the GMI is nonnegative and equals zero if and only if there are no statistical dependences. Unfortunately, the approach cannot be used for GCMI.

Note that the general conditional mutual information is an unbounded measure of conditional dependence. We use a transformed version of the mutual information and conditional mutual information

$$G(X,Y) = 1-\exp\{-I_q(X;Y)\}, \quad T(X;Y|Z) = 1-\exp\{-I_q(X;Y|Z)\}. \tag{9}$$

3. Test lags dependence using general mutual information measures

In this section, we propose to investigate and test for independence in time series using the general conditional mutual information measure defined above. The advantage of using

$T(X;Y|Z)$ as a test statistic is that it captures dependence while the conditioning on intermediate values of the time series will also give insights on the order of the underlying process. As a special case of $T(X;Y|Z)$, $G(X,Y)$ has the same estimation and asymptotic property.

For a stationary time series $\{X_t\}_{t=1}^n$ we define the M dimension delay vectors as $X_t^M = (X_{t-1},\ldots,X_{t-M})$, where M is a positive integer determined beforehand.

Let $X_t^{-j} = (X_{t-1},\ldots,X_{t-j+1},X_{t-j-1},\ldots,X_{t-M})$ denote the vector of the other variables in X_t^M except X_{t-j}. The null hypothesis is that

H$_0$: X_t and X_{t-j} are independent.

The general conditional mutual information between X_t and X_{t-j}, is

$$I_q(X_t;X_{t-j}|X_t^{-j}) = H_q(X_t,X_{t-j},X_t^{-j}) + H_q(X_t,X_t^{-j}) - H_q(X_{t-j},X_t^{-j}) - H_q(X_t^{-j}). \quad (10)$$

The statistics to test the dependence between X_t and X_{t-j} is

$$T_j(X_t;X_{t-j}|X_t^{-j}) = 1 - \exp\{-I_q(X_t;X_{t-j}|X_t^{-j})\}. \quad (11)$$

We use the generalized correlation integral to estimate the information quantities. The choice $q=2$ is by far the most popular in chaos analysis, since it allows for efficient estimation algorithms. From the results in [8], another reason for using $q=2$ is that, because of its symmetry, it is the fastest to compute all the generalized correlation integrals. The second order ($q=2$) correlation integral for the vector X is

$$C(X;\varepsilon) = \iint I_{(\|s-t\|\le\varepsilon)} f_X(s) f_X(t) \mathrm{d}s\mathrm{d}t, \quad (12)$$

where $I_{(\cdot)}$ denotes the indicator function taking values 0 and 1, and $\|\cdot\|$ denotes the supremum norm, $\|X\| = \sup_{i=1,\ldots,\dim X} |x_i|$. The parameter ε plays the role of a bandwidth.

Because (12) is just the expectation of the kernel function, it can be estimated straightforwardly in a U-statistics framework, by

$$\hat{C}(X;\varepsilon) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_{(\|X(i)-X(j)\|\le\varepsilon)}, \quad (13)$$

where n is the length of the observation series.

The relation of $H_2(X)$ and $C(X;\varepsilon)$ for ε small is

$$H_2(X) \simeq -\ln C(X;\varepsilon) + m\ln\varepsilon. \quad (14)$$

where m is the dimension of X.

Then the estimators for $G(X,Y)$ and $T(X,Y|Z)$ are

$$\hat{G}(X;Y) = 1 - \exp\{-\hat{I}_2(X;Y)\} = 1 - \frac{\hat{C}(X;\varepsilon)\hat{C}(Y;\varepsilon)}{\hat{C}(X,Y;\varepsilon)},$$

$$\hat{T}(X,Y|Z) = 1 - \exp\{-\hat{I}_2(X,Y|Z)\} = 1 - \frac{\hat{C}(X,Z;\varepsilon)\hat{C}(Y,Z;\varepsilon)}{\hat{C}(X,Y,Z;\varepsilon)\hat{C}(Z;\varepsilon)}. \quad (15)$$

Let $C^j(\varepsilon)$, $C_3^j(\varepsilon)$, $C_{21}^j(\varepsilon)$, $C_{22}^j(\varepsilon)$ be shorthand notation for $C(X_t^{-j};\varepsilon)$, $C(X_t,X_{t-j},X_t^{-j};\varepsilon)$,

$C(X_t, X_t^{-j}; \varepsilon)$, $C(X_{t-j}, X_t^{-j}; \varepsilon)$, respectively. Then the estimator $\hat{T}_j(\varepsilon)$ for T_j is

$$\hat{T}_j(\varepsilon) = 1 - \frac{\hat{C}_{21}^j(\varepsilon)\hat{C}_{22}^j(\varepsilon)}{\hat{C}_3^j(\varepsilon)\hat{C}^j(\varepsilon)}. \tag{16}$$

Theorem 4 *Under the null hypothesis of an i.i.d. process, the asymptotical distribution of $\hat{T}_j$ is*

$$n^{1/2}\hat{T}_j(\varepsilon) \xrightarrow{d} N(0, V_T(\varepsilon)) \tag{17}$$

where n is the series length, $\xrightarrow{d}$ denotes convergence in distribution. Since ε is fixed, we shall suppress ε in the notation,

$$C^j \equiv C^j(\varepsilon),\ C_3^j \equiv C_3^j(\varepsilon),\ C_{21}^j \equiv C_{21}^j(\varepsilon),\ C_{22}^j \equiv C_{22}^j(\varepsilon), \hat{T}_j \equiv \hat{T}_j(\varepsilon),\ V_T \equiv V_T(\varepsilon).$$

Proof Let $\{Y_t\}$ be an $\mathbb{R}^k$ -valued stochastic process. Then U-statistics of order 2 are defined by

$$U(n) = \frac{2}{n(n-1)}\sum_{s=1}^{n}\sum_{t=s}^{n} h(Y_t, Y_s), \tag{18}$$

where $h : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. Moreover, $h(y, z) = h(z, y)$ is a symmetric function and called a "kernel".

Since the indicator kernel in the correlation integral is bounded between 0 and 1, the moment conditions needed by Denker and Keller [9] are trivially satisfied. In addition, under the stationary and conditional independent assumptions the weak dependence conditions are also satisfied. The following (finite n) U-statistics and functions of U-statistics are defined.

$$g_j \equiv \hat{C}_{21}^j\hat{C}_{22}^j - \hat{C}_3^j\hat{C}^j \equiv G[\hat{C}_{21}^j, \hat{C}_{22}^j, \hat{C}_3^j, \hat{C}^j]. \tag{19}$$

Note that $\hat{T}_j \equiv D[\hat{C}_{21}^j, \hat{C}_{22}^j, \hat{C}_3^j, \hat{C}^j]$ and g_j are functions of U-statistics. Since $\hat{C}_{21}^j, \hat{C}_{22}^j, \hat{C}_3^j$ and $\hat{C}^j$ converge in probability to $C_{21}^j, C_{22}^j, C_3^j$ and C^j under the stationary and conditional independent assumptions as $n \to \infty$, respectively (The deduction is straightforward from the properties of U-statistics), it follows that g_j and $\hat{C}_j$ converge in probability to 0 as $n \to \infty$. To put it in another way, under i.i.d. D and G are 0 at population values.

Let D_k, G_k denote the partial derivatives with respect to the k-th argument of D and G respectively ($k = 1, 2, 3, 4$), evaluated at population values, C_{21}^j, C_{22}^j, C_3^j, C^j.

We expand the functions D and G around the population values in a Taylor series and take the limits as $n \to \infty$. This yields, for any smooth function H such that $H[C_{21}^j, C_{22}^j, C_3^j, C^j] = 0$,

$$\begin{aligned}
&\lim\{n^{1/2}H[\hat{C}_{21}^j, \hat{C}_{22}^j, \hat{C}_3^j, \hat{C}^j]\} \\
&\quad = \lim\{n^{1/2}[H_1 \times (\hat{C}_{21}^j - C_{21}^j) + H_2 \times (\hat{C}_{22}^j - C_{22}^j) + H_3 \times (\hat{C}_3^j - C_3^j) + H_4 \times (\hat{C}^j - C^j)]\} \\
&\quad = N(0, V(H)),
\end{aligned} \tag{20}$$

where all partial derivative H_i, $i = 1, 2, 3, 4$ are evaluated at $(C_{21}^j, C_{22}^j, C_3^j, C^j)$, "lim" denotes limit in distribution as $n \to \infty$ and $N(0, V)$ denotes Gaussian distribution with mean 0 and variance, V. Note that the right hand side of (20) is a linear combination of U-statistics, and hence a U-statistic for $H = G, D$. Let $\tilde{g}_j$ and $\tilde{T}_j$ denote the linear terms in (20) for $H = G, D$, respectively. In particular

$$\tilde{g}_j = C^j \times (\hat{C}_3^j - C_3^j) + C_3^j \times (\hat{C}^j - C^j) - C_{22}^j \times (\hat{C}_{21}^j - C_{21}^j) - C_{21}^j \times (\hat{C}_{22}^j - C_{22}^j), \tag{21}$$

$$\tilde{T}_j = \frac{\tilde{g}_j}{C_3^j C^j}. \tag{22}$$

Equ. (19) implies $\lim\{n^{1/2}\tilde{g}_j\} = N(0, V(G))$, $\lim\{n^{1/2}\tilde{T}_j\} = N(0, V(D))$, where

$$V(G) = \lim E[\{n^{1/2}\tilde{g}_j\}^2], \quad V(D) = \lim E[\{n^{1/2}\tilde{T}_j\}^2]. \tag{23}$$

The limits in (22) can be calculated by methods similar to Wu, Savit and Brock [10]. After complex calculations, the variance is

$$V_T = 4\sigma^2(\hat{T}_j) = 4\frac{\sigma^2(g_j)}{(C_3^j)^2(C^j)^2},$$

where

$$\sigma^2(g_j) = (C_3^j)^2 \mathrm{Var}(K_1^C) + (C^j)^2 \mathrm{Var}(K_{1,ab}^C) - (C_{22}^j)^2 \mathrm{Var}(K_{1,b}^C) - (C_{21}^j)^2 \mathrm{Var}(K_{1,a}^C),$$

$$K^C \equiv K^C[Z(s), Z(t)] = I_{(\|Z(s)-Z(t)\| \le \varepsilon)},$$

$$K_{ab}^C \equiv K^C[Z_{ab}(s), Z_{ab}(t)] = I_{(\|Z_{ab}(s)-Z_{ab}(t)\| \le \varepsilon)},$$

$$K_1^C = \int K^C[Z(s+i), Z(t+i)]\mathrm{d}F[Z(s+i)],$$

$$K_{1,ab}^C = \int K_{ab}^C[Z_{ab}(s+i), Z_{ab}(t+i)]\mathrm{d}F[Z_{ab}(s+i)],$$

$$Z_{ab}(t) = (X_t, X_{t-j}, X_t^{-j}), Z(t) = (X_t^{-j}), Z_a(t) = (X_t, X_t^{-j}), Z_b(t) = (X_{t-j}, X_t^{-j}). \qquad \square$$

In practical applications, we should pay attention to the choice of the parameter ε. From (14) and the proof process of Theorem 4, the accuracy of the estimator increases when ε decreases, while the standard deviation decreases when ε decreases. In general, one selects $\varepsilon = 0.5\sigma_X, 1.0\sigma_X, 1.5\sigma_X$, where σ_X is the standard deviation of the process.

The test procedure is composed of the following steps:

Step 1. For chosen significance level, calculate the critical values C_s for observation series $\{X_t\}_{t=1}^n$. We suggest a simulation algorithm instead of calculating the variance in Theorem 4 straightly.

Step 2. Calculate the test statistic $T_j(X_t; X_{t-j}|X_t^{-j})$.

Step 3. Reject the independence of X_t and X_{t-j}, if $T_j(X_t; X_{t-j}|X_t^{-j}) > C_s$.

4. Simulation results

To investigate the potential applicability of the $\hat{G}$ defined in the previous section some simulation experiments are presented. Unless stated otherwise, the replication number for all simulations is 200. It has been found that $\hat{G}$ is biased in finite samples. In order to show the relation between the bias and the sample size, a set of Gaussian independent random variables y_t with samples sizes of 100,200,300,500,1000 and 3000 are generated and $\hat{G}_j = \hat{G}(y_t, y_{t-j})$ for $j = 1, \ldots, 10$ are computed. The means and standard deviation are tabulated in Table 1.

Lag	100	200	300	500	1000	3000
1	-0.0077	-0.0035	-0.0024	0.0005	0.0003	0.0374×10^{-3}
	(0.0439)	(0.0266)	(0.0191)	(0.0129)	(0.0098)	(0.0050)
2	-0.0095	-0.0038	-0.0011	-0.0009	-0.0013	0.2572×10^{-3}
	(0.0471)	(0.0249)	(0.0178)	(0.0147)	(0.0097)	(0.0051)
3	-0.0035	-0.0019	-0.0009	-0.0006	-0.0006	-0.8706×10^{-3}
	(0.0446)	(0.0256)	(0.0196)	(0.0162)	(0.0095)	(0.0050)
4	-0.0041	-0.0039	0.0006	-0.0026	-0.0009	0.0106×10^{-3}
	(0.0485)	(0.0254)	(0.0199)	(0.0149)	(0.0095)	(0.0055)
5	-0.0036	-0.0009	-0.0014	-0.0000	-0.0011	-0.0435×10^{-3}
	(0.0486)	(0.0270)	(0.0190)	(0.0137)	(0.0093)	(0.0052)
6	0.0013	-0.0014	-0.0007			
	(0.0474)	(0.0268)	(0.0202)			
7	-0.0073	-0.0025	-0.0027			
	(0.0544)	(0.0234)	(0.0191)			
8	-0.0094	-0.0036	-0.0012			
	(0.0511)	(0.0288)	(0.0188)			
9	-0.0041	0.0002	-0.0019			
	(0.0481)	(0.0278)	(0.0206)			
10	-0.0007	-0.0048	-0.0021			
	(0.0503)	(0.0239)	(0.0190)			

Table 1 The mean and standard deviation of $\hat{G}_j$ for various sample sizes

For simplicity, we use the same ε for all the models. The scale parameter is taken to be $\varepsilon = 0.5$ (after rescaling each time series to zero mean and unit variance). The reason is that, when $\varepsilon = 0.5$, the standard deviation of $\hat{G}$ is the nearest to that in [1].

Lag	Mean	Standard deviation	90%	95%	99%
1	-0.0010	0.0200	0.0232	0.0344	0.0453
2	-0.0007	0.0201	0.0252	0.0332	0.0501
3	0.0007	0.0206	0.0265	0.0343	0.0500
4	-0.0007	0.0194	0.0258	0.0334	0.0471
5	-0.0010	0.0196	0.0252	0.0340	0.0481
	0	0.02	0.0256	0.0328	0.0466

Table 2 Mean, standard deviation and critical values for $\hat{G}_j$

The mean, standard deviation and empirical critical values of G were estimated for sample size 300 using 1000 replications and are reported in Table 2. For this sample size, the appropriate 95% critical value seems to be 0.03 for checking the null hypothesis of independence. The last row is the corresponding critical values for Gauss distribution $N(0, 0.02^2)$. The values are very close to those of simulation, which verify Theorem 4 in a simulation way.

Various time series were generated to examine the performance of $\hat{G}$. For the purpose of comparison, we use the same models considered by Granger and Lin [1] too, which are listed below, with $e_t \sim$ i.i.d, $N(0,1)$.

Model 1 $y_t = e_t + 0.8e_{t-1}^2$;
Model 2 $y_t = e_t + 0.8e_{t-2}^2$;
Model 3 $y_t = e_t + 0.8e_{t-3}^2$;
Model 4 $y_t = e_t + 0.8e_{t-1}^2 + 0.8e_{t-2}^2 + 0.8e_{t-3}^2$;
Model 5 $y_t = |y_{t-1}|^{0.8} + e_t$;
Model 6 $y_t = \text{sign}(y_{t-1}) + e_t$;
Model 7 $y_t = 0.8y_{t-1} + e_t$;
Model 8 $y_t = y_{t-1} + e_t$;
Model 9 $y_t = 0.6e_{t-1}y_{t-2} + e_t$;
Model 10 $y_t = 4y_{t-1}(1 - y_{t-1}) + e_t$ for $t > 1, 0 < y_1 < 1$.

Lag	Model 1	Model 2	Model 3	Model 4
1	0.1267	0.0009	-0.0028	0.2138
	(0.0254)	(0.0270)	(0.0247)	(0.0322)
2	-0.0008	0.1214	-0.0004	0.1032
	(0.0231)	(0.0252)	(0.0221)	(0.0293)
3	-0.0032	-0.0062	0.1230	0.0369
	(0.0234)	(0.0209)	(0.0264)	(0.0258)
4	-0.0012	-0.0013	0.0003	0.0018
	(0.0242)	(0.0251)	(0.0230)	(0.0261)
5	0.0001	-0.0051	-0.0024	0.0012
	(0.0238)	(0.0227)	(0.0245)	(0.0264)
6	-0.0014	-0.0050	-0.0012	-0.0011
	(0.0235)	(0.0236)	(0.0238)	(0.0270)
7	-0.0031	-0.0037	-0.0032	0.0006
	(0.0242)	(0.0247)	(0.0227)	(0.0267)
8	-0.0027	-0.0032	-0.0026	0.0011
	(0.0256)	(0.0225)	(0.0243)	(0.0265)
9	-0.0034	0.0012	-0.0018	0.0009
	(0.0231)	(0.0265)	(0.0249)	(0.0264)
10	-0.0007	-0.0036	-0.0041	-0.0014
	(0.0230)	(0.0210)	(0.0233)	(0.0249)

Table 3 Mean and standard deviation of $\hat{G}_j$ for nonlinear MA models

The results are listed in Table 3. Model 1 is a nonlinear MA(1) so that, theoretically, all G_i except G_1 should be zero. Column 1 in Table 3 displays exactly this pattern. $\hat{G}_1$ is 0.12 and all other $\hat{G}_i$'s are close to or less than 0, which is the average value for the independent case. Similar results hold for models 2 and 3, which are nonlinear MA(2) and MA(3) with middle coefficients set to zero. Model 4 is again nonlinear MA(3) but with more nonzero coefficients. Using the critical values in Table 2, $\hat{G}_1$, $\hat{G}_2$ and $\hat{G}_3$ in column 4 of Table 3 are all significant.

The simulation results for the later 6 models are listed in Table 4. For the autoregressive models 5, 6 and 7, the $\hat{G}_i$ decreases smoothly as i increases. Finally, consider the random-walk time series in model 8. The $\hat{G}_i$ remains at a fairly high level and decreases very slowly with i, which is very similar to the linear case. Model 9 is a bilinear model. $\hat{G}_1$ and $\hat{G}_2$ are significant, which demonstrates again the ability of G to detect the nonlinear structure of a time series. Model 10 is a chaotic time series, the $\hat{G}_1$ and $\hat{G}_2$ are significant.

Lag	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
1	0.1822	0.2447	0.3534	0.6359	0.0501	0.5632
	(0.0367)	(0.0322)	(0.0418)	(0.0386)	(0.0327)	(0.0064)
2	0.0576	0.1262	0.2063	0.5999	0.0741	0.2024
	(0.0335)	(0.0373)	(0.0458)	(0.0515)	(0.0295)	(0.0239)
3	0.0201	0.0601	0.1296	0.5695	0.0120	-0.0378
	(0.0257)	(0.0322)	(0.0474)	(0.0640)	(0.0256)	(0.0200)
4	0.0088	0.0294	0.0831	0.5437	0.0179	-0.0093
	(0.0231)	(0.0276)	(0.0469)	(0.0740)	(0.0246)	(0.0177)
5	0.0020	0.0138	0.0550	0.5206	0.0005	-0.0212
	(0.0228)	(0.0241)	(0.0422)	(0.0827)	(0.0226)	(0.0141)
6	0.0003	0.0078	0.0379	0.5000	0.0053	-0.0055
	(0.0223)	(0.0212)	(0.0378)	(0.0900)	(0.0226)	(0.0132)
7	0.0003	0.0033	0.0263	0.4811	-0.0021	0.0002
	(0.0202)	(0.0202)	(0.0346)	(0.0966)	(0.0221)	(0.0136)
8	0.0002	0.0016	0.0180	0.4632	0.0016	-0.0011
	(0.0215)	(0.0181)	(0.0324)	(0.1022)	(0.0224)	(0.0130)
9	0.0000	-0.0001	0.0144	0.4471	-0.0011	-0.0008
	(0.0231)	(0.0190)	(0.0310)	(0.1070)	(0.0223)	(0.0133)
10	0.0025	0.0008	0.0125	0.4324	-0.0020	-0.0008
	(0.0214)	(0.0198)	(0.0302)	(0.1110)	(0.0221)	(0.0140)

Table 4 Mean and standard deviation of $\hat{G}_j$ for nonlinear AR models

Except for Model 10, we get the same results with the statistic R in [1], which verify that I_2 behaves much like I_1 for those models.

In order to identify exact lags for the autoregressive models 5, 6, 7 and 8, the appropriate statistics should be the conditional mutual measure $\hat{T}$. Thus a very large amount of data is required to achieve reasonable accuracy. The scale parameter is taken to be $\varepsilon = 1.0$ to capture more information and $M = 5$ in the delay vectors. The mean, standard deviation and empirical critical values of $\hat{T}$ were estimated for sample size 500 using 1000 replications and are listed in

Table 5. The last row is the corresponding critical values for Gauss distribution $N(0, 0.0145^2)$. The values verify Theorem 4 in a simulation way.

Lag	Mean	Standard deviation	90%	95%	99%
1	-0.0011	0.0145	0.0178	0.0226	0.0340
2	-0.0010	0.0154	0.0182	0.0236	0.0355
3	-0.0008	0.0144	0.0179	0.0225	0.0333
4	-0.0008	0.0144	0.0168	0.0226	0.0314
5	-0.0009	0.0144	0.0176	0.0238	0.0314
	0	0.0145	0.0185	0.0237	0.0337

Table 5 Mean, standard deviation and critical values for $\hat{T}_j$

The results are reported in Table 6. For models 5, 6 and 7, all the $\hat{T}_i$ are not significant except for $\hat{T}_1$, which is exactly as expected. For the random-walk time series in model 8, only $\hat{T}_1$ is significant too. However, the Kendall's partial τ_i are all significant from lag 1 up to lag 8 in [1]. In that case the statistic T behaves better than τ.

Lag	Model 5	Model 6	Model 7	Model 8	Model 10
1	0.0877	0.1002	0.1204	0.0250	0.2166
	(0.0134)	(0.0132)	(0.0107)	(0.0145)	(0.0087)
2	0.0012	0.0152	0.0034	0.0031	-0.0305
	(0.0105)	(0.0088)	(0.0056)	(0.0016)	(0.0117)
3	-0.0003	0.0016	0.0003	0.0015	0.0155
	(0.0091)	(0.0086)	(0.0050)	(0.0007)	(0.0108)
4	-0.0001	0.0003	0.0002	0.0010	-0.0142
	(0.0091)	(0.0092)	(0.0042)	(0.0005)	(0.0081)
5	0.0000	-0.0007	0.0010	0.0018	-0.0011
	(0.0105)	(0.0094)	(0.0053)	(0.0009)	(0.0082)

Table 6 Mean and standard deviation of $\hat{T}_j$ for nonlinear AR models

Furthermore, we introduce another four models to verify the power of our statistics T_j.

Model 11 $y_t = \begin{cases} 0.7y_{t-1} + e_{1t} & y_{t-2} < 0 \\ 0.5y_{t-3} + e_{2t} & y_{t-2} \geq 0 \end{cases}, \quad e_{1t} \sim N(0, 0.25), \quad e_{2t} \sim N(0, 1);$

Model 12 $y_t = \sigma_t e_t, \quad \sigma_t^2 = 0.3y_{t-1}^2 + 0.5\sigma_{t-1}^2, \quad e_t \sim N(0, 1);$

Model 13 $y_t = (0.3 + 0.9y_{t-1})e^{-3y_{t-1}^2}, +e_t, \quad e_t \sim N(0, 0.25);$

Model 14 $y_t = \sin(y_{t-2})y_{t-1} + e_t, \quad e_t \sim N(0, 1).$

The results are presented in Table 7. For the SETAR model 11, the EXPAR model 13 and the FAR model 14, the statistic T_n gives correct lag dependence. In the GARCH(1,1) model 12, $\hat{T}_1$, $\hat{T}_2$, $\hat{T}_3$ are all significant, this is because that the dependence is caused by latent variable.

Lag	Model 11	Model 12	Model 13	Model 14
1	0.0908	0.0498	0.0446	0.0862
	(0.0131)	(0.0129)	(0.0152)	(0.0144)
2	0.0360	0.0352	0.0001	0.0390
	(0.0088)	(0.0119)	(0.0140)	(0.0097)
3	0.0259	0.0270	-0.0004	0.0014
	(0.0128)	(0.0112)	(0.0123)	(0.0082)
4	0.0074	0.0191	0.0000	-0.0018
	(0.0080)	(0.0118)	(0.0129)	(0.0083)
5	0.0013	0.0157	-0.0023	-0.0010
	(0.0097)	(0.0122)	(0.0133)	(0.0092)

Table 7 Mean and standard deviation of $\hat{T}_j$ for new models

5. Conclusion

In this paper, we consider statistics based on GMI and GCMI for identifying what lags to use in a nonlinear model. The statistics are invariant with one-to-one transformation. In the case of multivariate Gaussian distribution, $T(X;Y|Z)$ is strictly increasing function of $|\rho(X;Y|Z)|$, where $\rho(X;Y|Z)$ is usual partial correlation coefficient between X and Y given Z. The correlation integral estimates are faster to compute than the kernel density estimates and have Gaussian distributions under the independence hypothesis. The same simulation results between $G(X,Y)$ and $R(X,Y)$ in [1] shows the availability of the 2-order GMI for testing independence. In general the simulations produce satisfactory results in identifying the correct lags of the true models.

References

[1] GRANGER C, LIN J L. *Using the mutual information coefficient to identify lags in nonlinear models* [J]. J. Time Ser. Anal., 1994, **15**(4): 371–384.

[2] GRANGER C W, MAASOUMI E, RACINE J. *A dependence metric for possibly nonlinear processes* [J]. J. Time Ser. Anal., 2004, **25**(5): 649–669.

[3] MAASOUMI E, RACINE J. *Entropy and predictability of stock market returns* [J]. J. Econometrics, 2002, **107**(1-2): 291–312.

[4] DARBELLAY G, WUERTZ D. *The Entropy as a Tool for Analysing Statistical Dependence's in Financial Time Series* [J]. Phys. Lett. A, 2000, **287**: 429–439.

[5] DIKS C G H, MANZAN S. *Test for serial independence and linearity based on correlation integrals* [J]. Studies in Nonlinear Dynamics & Econometrics, 2002, **6**: 1–20.

[6] WHITTAKER J. *Graphical Models in Applied Multivariate Statistics* [M]. John Wiley, Chichester, 1990.

[7] POMPE B. *Measuring statistical dependences in a time series* [J]. J. Statist. Phys., 1993, **73**(3-4): 587–610.

[8] GRASSBERGER P. *Finite sample corrections to entropy and dimension estimates* [J]. Phys. Lett. A, 1988, **128**(6-7): 369–373.

[9] DENKER M, KELLER G. *On U-statistics and v. Mises' statistics for weakly dependent processes* [J]. Z. Wahrsch. Verw. Gebiete, 1983, **64**(4): 505–522.

[10] WU K, SAVIT R, BROCK W. *Statistical tests for deterministic effects in broad band time series* [J]. Phys. Lett. D, 1993, **69**: 172–188.