韓國水資源學會論文集
第30卷 第6號 · 1997年 12月
pp. 641~648

# 시계열 수문자료의 비선형 상관관계
# How to Measure Nonlinear Dependence in Hydrologic Time Series

문 영 일*
Moon, Young-Il

## Abstract

Mutual information is useful for analyzing nonlinear dependence in time series in much the same way as correlation is used to characterize linear dependence. We use multivariate kernel density estimators for the estimation of mutual information at different time lags for single and multiple time series. This approach is tested on a variety of hydrologic data sets, and suggested an appropriate delay time $\tau$ at which the mutual information is almost zero then multi-dimensional phase portraits could be constructed from measurements of a single scalar time series.

*Keywords* : mutual information, correlation, delay time, kernel density estimators

## 요 지

상관계수가 변수간의 선형 상관관계를 나타내듯이 mutual information은 변수간의 비선형 상관관계를 나타내 준다. 본 논문에서는 mutual information 추정법으로 다변수 핵 밀도함수 (multivariate kernel density estimator)를 이용한 방법이 여러 time lags값에 대하여 산정 되었다. 많은 수문자료에서 보여지는 비선형 관계를 Mutual Information으로 확인하여 보았고, 또한 Mutual Information값이 거의 0인 점에서 optimal delay time을 구하여, 하나의 자료로부터 다변수 회귀분석 모델을 만들 때 이용할 수 있다.

**핵심용어**: mutual information, 상관계수, 지체시간, 핵 밀도함수

* 서울시립대학교 토목공학과 전임강사

## 1. Introduction

It is common to find in a time series of hydrologic data that an observation at one time period is strongly dependant with the observation in the preceding time period. Correlation function is frequently used to quantity this dependence. The correlation function hitherto measures only the linear dependence, which may be sufficient in most situations to explain the dependence, but in general it is desirable to consider also nonlinear relationships between different variables. Given that there are feedbacks and interactions between hydrologic processes it is of interest to look for a measure of nonlinear dependence. The motivation for considering the mutual information is its capability to measure a general dependence between two variables. If the two variables are independent then the mutual information between them is zero. However, if the two variables are strongly dependent then the mutual information between them is large. The mutual information measures the general dependence of two variables while the correlation function measures the linear dependence. For example, there is a strong evidence of a nonlinear association between nutrient level and the number of fish in Figure 1. Note that the strength of the linear relationship is almost zero (i. e. r2=0), but the mutual information shows a strong relationship between the variables. Therefore, mutual information provides a better criterion for the measure of the dependence between variables than the correlation function.. A detailed investigation of the advantages of the mutual information versus the correlation function is contained in Li(1990).

Another objective of mutual information (M.I.) analysis is to measure how dependent the values of $x(t+\tau)$ are on the values of $x(t)$ where $\tau$ is a delay time. There has been a growing interest in phase-portrait reconstruction from time series data in fields as diverse as hydrology (Moon and Lall, 1996; Abarbanel et al., 1996) and hydrodynamics (Brandstater et al., 1983). If we can get an appropriate delay time $\tau$ at which the mutual information is almost zero then multi-dimensional phase portraits could be constructed from measurements of a single scalar time series. In this approach portraits are constructed by expanding a scalar time series $x(t)$ into a vector time series $X(t)$ using time delays $\tau$: $X(t) = \{x_1(t), x_2(t), x_3(t), \ldots, xM(t)\}$, where $xM(t) = x(t+M\tau)$. If the delay time is too small, the reconstructed attractor is restricted to the diagonal of the

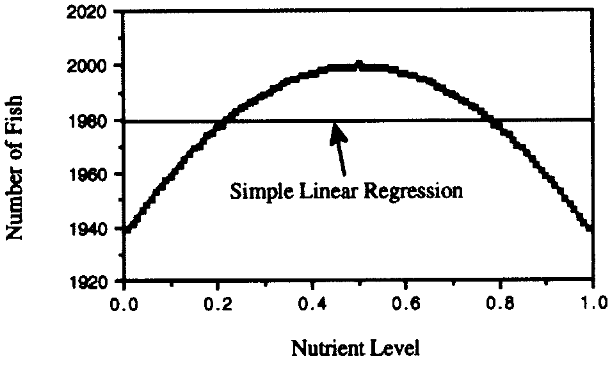

Fig. 1. Data on Fish Population vs. Nutrients

reconstruction space because $x(t)$ and $x(t+\tau)$ will basically be the same. On the other hand if $\tau$ is chosen too large then the attractor coordinates are uncorrelated and the system is chaotic. Thus, all relevant information for phase space reconstruction is lost since neighboring trajectories diverge, and averaging in time and/or space is no longer useful.

No criteria for choosing $\tau$ exists in literature until Fraser and Swinney (1986) proposed the use of mutual information (M.I.) as a criterion for choosing $\tau$ and argue that this provides an excellent criterion for choosing $\tau$ in most systems. They suggest that value of $\tau$ produce the first local minimum of mutual information. This choice is better than choosing $\tau$ as the lag at which autocorrelation function (ACF) first passes through zero, as the ACF only measures the linear dependence, while the M.I. measures the general dependence of two variables and hence provides a better criterion (Graf and Elbert, 1990) for the choice of $\tau$.

Fraser and Swinney (1986) developed the use of multivariate histogram for the estimation of M.I. and subsequent choosing of $\tau$. Here we propose the use of nonparametric multivariate kernel density estimator for the estimation of M.I. Our investigations show that this is particularly advantageous with small data sets.

## 2. Definition of the Mutual Information

Mutual Information (Fraser and Swinney, 1986) provides a general measure of dependence between two variables. Let us denote the time series of the two variables as $s_1$, $s_2$, ...$s_i$, ...$s_n$, and $q_1$, $q_2$, ...$q_j$, ...$q_n$, where $n$ is the record length, and the sampling rate $\partial t$ is fixed. The mutual information between observations $s_i$ and $q_j$ is defined in bits as:

$$MI_{s,q}(s_i,q_j)=\log_2\left(\frac{P_{s,q}(s_i,q_j)}{P_s(s_i)P_q(q_j)}\right) \qquad (1)$$

where $P_{s,q}(s_i,q_j)$ is the joint probability density of $s$ and $q$ evaluated at $(s_i, q_j)$, and $P_s(s_i)$ and $P_q(q_j)$ are the marginal probability densities of s and $q$ evaluated at $s_i$ and $q_j$ respectively.

Where overall dependence between the two series is of interest, one can define (analogously to linear correlation) the Average Mutual Information $I_{s,q}$ as:

$$I_{s,q}=\sum_{i,j}P_{s,q}(s_i,q_j)\log_2\left(\frac{P_{s,q}(s_i,q_j)}{P_s(s_i)P_q(q_j)}\right) \qquad (2)$$

This measure is useful for identifying components in multivariate sampling that seem to be related or independent. A particular recent use (Martinerie et al., 1992; Abarbanel, 1994; Gao, 1994) is the choice of an appropriate delay parameter while reconstructing a state space from an experimental time series.

Kernel density estimation is a nonparametric method for estimating probability densities. We learn from the statistical literature (Silverman, 1986, Devroye and Györfi, 1985; Scott, 1992) that kernel density estimates can be superior to the histogram in terms of (1) better Mean Square Error rate of convergence of the estimate to the underlying density, (2) insensitivity to the choice of origin, and (3) ability to specify more sophisticated window shapes than the rectangular window for "binning" or frequency counting.

A kernel density estimate (KDE) of a vector $y$ is given (Silverman, 1986) as:

$$(y)=\frac{1}{n}\sum_{j=1}^{n}K(u) \qquad (3)$$

where

$$u=\frac{(y-y_i)^T S^{-1}(y-y_i)}{h^2} \qquad (4)$$

where, $K(u)$ is a multivariate kernel function., $y=[y_1,y_2,\ldots,y_d]^T$ is the $d$ dimensional random vector whose density is being estimated; $y_i=[y_{1i},y_{2i},\ldots,y_{di}]^T$, $i=1$ to $n$ are the n sample vectors, $h$ is the kernel bandwidth and $S$ is the covariance matrix of the $y_i$. The kernel

Table 1. Description of Data Sets Used.

| | |
|---|---|
| Data from AR(1) model | 500 data points were generated from the AR (1) model: $x_t = \rho x_{t-1} + \sqrt{1-\rho^2} N(1,0)$ where N(0,1) refers to a standard Gaussian density and $\rho = 0.85$. |
| GSL Monthly Volume data | Monthly volume of Great Salt Lake (GSL) for the period from Nov. 1847 to Dec. 1996. |
| Southern Oscillation Index (SOI) | Monthly mean difference in Sea Level Pressure (SLP) at Tahiti and Darwin from Sep. 1932 to Nov. 1993, 735 data points. SOI = SLP(Tahiti)-SLP(Darwin) |

function $K(u)$ is required to be a valid probability density function. In this paper we use the multivariate Gaussian probability density function for $K(u)$ which is given as,

$$K(u) = \frac{1}{(2\pi)^{d/2} h^d \, det(S)^{1/2}} \, exp(-u/2) \qquad (5)$$

An evaluation of $K(u)$ represents the weight given to an observation $y_i$, that is based on distance between $y$ and $y_i$. The distance used here is the Euclidean distance modified to recognize the covariance in the co-ordinates. We can see from (3) that the kernel estimator is a local weighted average of the relative frequency of observations in the neighborhood of the point of estimate. The kernel function, $K(.)$ prescribes the relative weights, and $h$ prescribes the range of data values over which the average is computed. The role of the covariance matrix $S$ is to recognize possible linear dependence amongst the coordinates. Its use allows one to appropriately orient the resulting kernel function and vary the bin width in proportion to the scale of variation in the rotated coordinates.

There are many methods for choosing the bandwidth $h$. Some of the best ones in the statistical literature are due to Sheather and Jones (1991), for $d=1$, and Wand and Jones (1994) for $d=2$. The computational burden associated with these and other data driven, automatic bandwidth selectors can be formidable. Here we made an expedient choice of the bandwidth as the one that minimizes the mean integrated square error (MISE) in ($y$) if the underlying distribution is assumed to be multivariate Gaussian. While this is not a theoretically satisfying choice, its performance in our tests was comparable, and computation time was orders of magnitude lower than the more rigorous choices. The "optimal" Gaussian bandwidth corresponding to the kernel choice in (5), is given by Silverman (1986) as:

$$h = \left[\frac{4}{(d+2)}\right]^{1/(d+4)} n^{-1/(d+4)} \qquad (6)$$

## 3. Data Sets

In order to demonstrate the application of the KDE to estimation of M.I. and the subsequent picking of the optimal delay time τ, one simulated time series and two real time series are chosen. The details of the data sets are given in Table 1.

## 4. Results and Conclusion

The mutual information is calculated for up to lags 100 for each of the data sets using the KDE and ACF up to 100 lags is also calculated for the data sets. Moon et al. (1995) estimated the M.I. for several simulated data sets using the histogram method (FSH) of Fraser and Swinney (1986) for comparison with the KDE approach. They represent that KDE provides an attractive alternative to the FSH method for estimating the average sample mutual information. Our results are consistent with these reported in Moon et al. (1995). For selected cases, it was possible to analytically compute the requisite probabilities and use

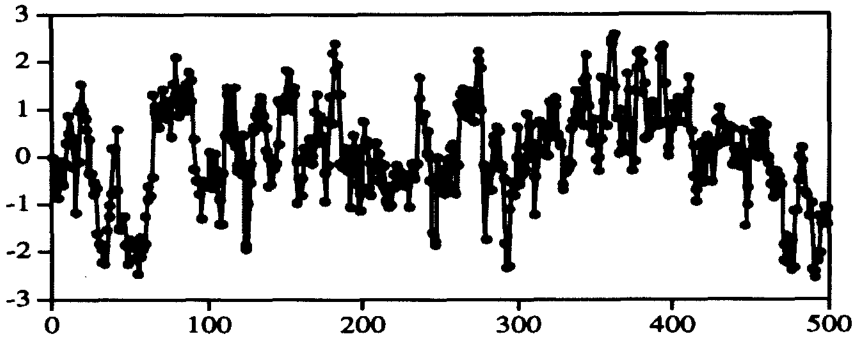

Fig. 2(a). AR(1) data

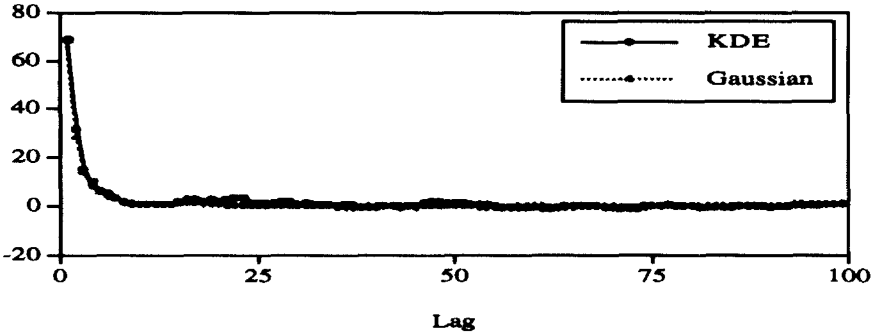

Fig. 2(b). $I_{x_t,x_{t-\tau}}$ from KDE and from fitted Gaussian densities for the AR(1) data.

them to derive the expected sample estimates of $I_{x_t},x_{t-\tau}$. In these cases, we found that the KDE estimates were numerically quite close to those from the analytical expressions.

The results for the data from AR(1) model is shown in Figure 2. Note that for an AR model the joint and marginal densities, $P_{x_t},x_{t-\tau}(\cdot)$, $P_{x_t}(\cdot)$, and $P_{x_{t-\tau}}(\cdot)$respectively are all Gaussians and hence $I_{x_t},x_{t-\tau}$ can be calculated directly by fitting Gaussian distributions to the data. From Figure 2, we observe that there is little difference in the analytical and KDE estimates of $I_{x_t},x_{t-\tau}$ The lag $\tau^*$ would be selected as 11 from KDE and from the analytical expression.

In Figure 3(a) and (b), GSL data and ACF of GSL monthly volume data is shown. The M.I. for KDE suggests a lag of 7 for $\tau$ in Figure 3(b). The next data set we considered Southern Oscillation Index (SOI). Figure 4(a) presents the data of Southern Oscillation Index(SOI) time series. In Figure 4(b), ACF of Southern Oscillation Index (SOI) is shown. The mutual information of KDE shows that the first minimum is at the lag of 11 months in Figure 4(c).

The purpose of the experiment was to test the multivariate kernel density estimator (KDE) for picking the optimal delay time $\tau$ and to compare its performance with Fraser and Swenney′ histogram (FSH) (1986). The mutual information of Fraser and Swenney histogram (FSH) dose not seem consistent (Moon et al., 1995). It may be from the histogram drawback about the choice of bin

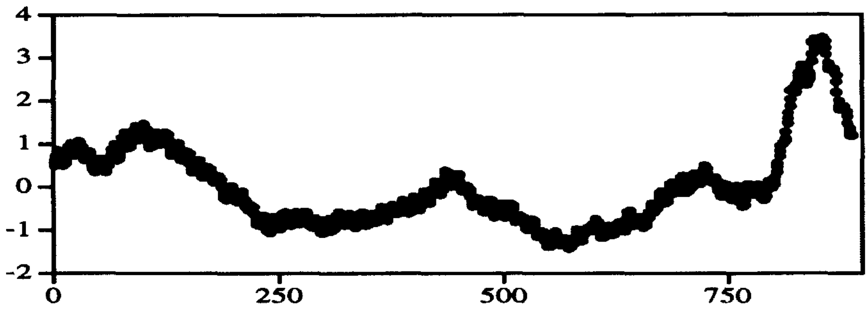

Fig. 3(a). Great Salt Lake monthly volume time series.

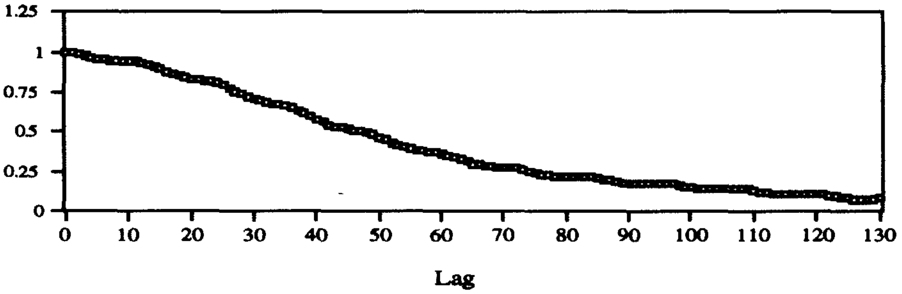

Fig. 3(b). ACF of Great Salt Lake monthly volume time series.

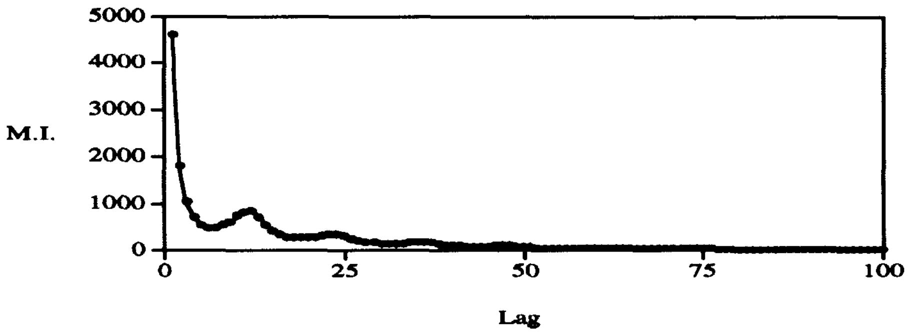

Fig. 3(c). $I_{x_t, x_{t-\tau}}$ of KDE for the Great Salt Lake monthly volume time series.

width which, primarily, controls the amount of smoothing inherent in the procedure. The usefulness of the nonparametric multivariate kernel density estimator in analyzing the mutual information is shown. The nonparametric multivariate kernel density estimator (KDE) provides more reliable mutual information.

Another purpose for this work was to investigate the optimum delay time $\tau$ for nonlinear hydrologic systems. If we know an appropriate $\tau$ then multidimensional phase portraits can be constructed from a single scalar time series.

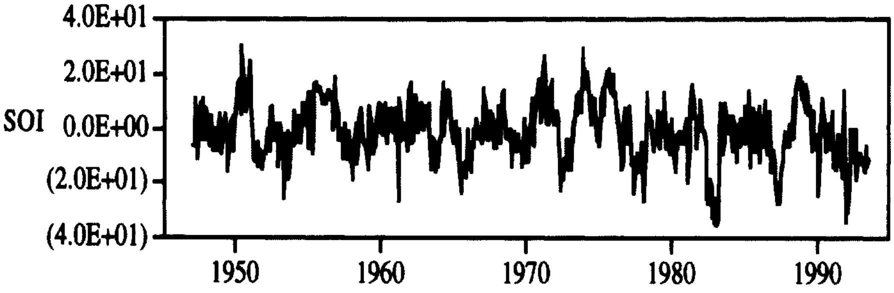

Fig. 4(a). Southern Oscillation Index(SOI) time series.

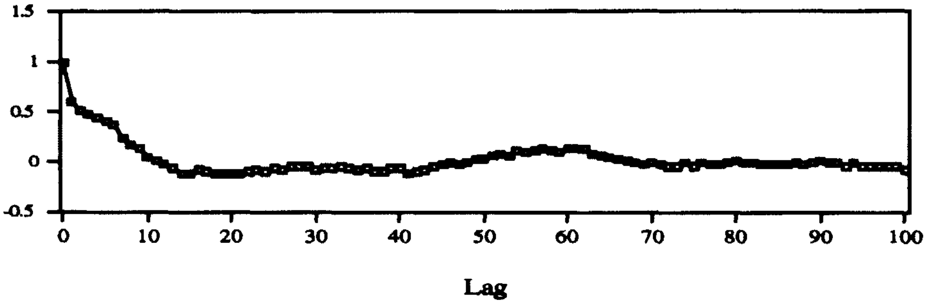

Fig. 4(b). ACF of Southern Oscillation Index(SOI)

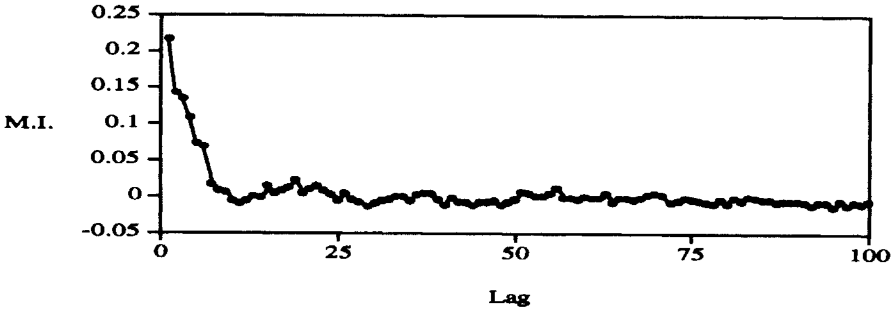

Fig. 4(c). $I_{x_t, x_{t-\tau}}$ of KDE for the Southern Oscillation Index (SOI).

## Acknowledgements

Partial support of this work by the USGS grant # 1434-92-G-226 and NSF grant # EAR-9205727 is acknowledged.

## 5. References

Abarbanel, H. D. I., Carroll, T.A., Pecora, L.M., Sidorowich, J.J., and Tsimring, L.S. (1994). "Predicting physical variables in time-delay embedding." *Physical Review E*, Vol. 49, pp. 1840-1853.

Abarbanel, H. D. I., Lall, U., Moon, Young-Il, Mann, M., and Sangoyomi, T. (1996). "Nonlinear-Dynamics and the Great Salt Lake: a Predictable Indicator of Regional

Climate." *Energy*, Vol. 21(7/8), pp. 655-665.

Brandstater, A., Swift, J., Swinney, H.L., Wolf, A., Farmer, D., Jen, E., and Crutchfield, J. (1983). "Low-dimensional chaos in hydrodynamic system." *Phys. Rev. Lett.*, Vol. 51, pp. 1442-1445.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L1 View*, John Wiley, New York.

Fraser, A.M. and Swinney, H.L. (1986). " Independent coordinates for strange attractors from mutual information." *Physical Review A*, Vol. 33, pp. 1134-1140.

Gao, J. and Zheng, Z. (1994). "Direct dynamical test for deterministic chaos and optimal embedding of a chaotic time series." *Physical Review E*, Vol. 49, pp. 3807-3814.

Graf, K. E. and Elbert, T. (1990). *Dimensional analysis of the waking EEG. Chaos in brain function*, Springer-Verlag, Erol Basar.

Li, W. (1990). "Mutual information functions versus correlation function." *Journal of Statistical Physics*, Vol. 60, pp. 823-837.

Martinerie, J. M., Albano, A. M., Mees, A. I., and Rapp, P. E. (1992). "Mutual information, strange attractors, and the optimal estimation of dimension." *Physical Review A* , Vol. 45, pp. 7058-7064.

Moon, Young-Il, Rajagopalan, B., and Lall, U. (1995). "Estimation of mutual information using kernel density estimators." *Physical Review Latter E*, Vol. 52, pp. 2318-2321.

Moon, Young-Il and Lall, U. (1996). "Atmospheric flow indices and interannual Great Salt Lake variability." *Journal of Hydrologic Engineering, ASCE*, Vol. 1, No. 2, pp. 55-62.

Scott, D.W. (1992). *Multivariate density estimation*, John Wiley and Sons, New York.

Sheather, S. J. and Jones, M. C. (1991). "A reliable data-based bandwidth selection method for kernel density estimation." *Journal of the Royal Statistical Society B*, Vol. 53, pp. 683-690.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall, New York.

Wand, M. P. and Jones, M. C. (1994). "Comparison of smoothing parameterizations in bivariate kernel density estimation." *Computational Statistics*, Vol. 9, pp. 97.

〈최종본 접수일 : 1997년 10월 13일〉