Second-order moments and mutual information in the analysis of time series

David R. Brillinger *
Statistics Department
University of California
Berkeley, CA, 94720-3860
E-mail: brill@stat.berkeley.edu

Abstract

A statistical network is a collection of nodes representing random variables and a set of edges that connect the nodes. A probabilistic model for such is called a statistical graphical model. These models, graphs and networks are particularly useful for examining statistical dependencies amongst quantities via conditioning. In this article the nodal random variables are time series. Basic to the study of statistical networks is some measure of the strength of (possibly directed) connections between the nodes. The use of the ordinary and partial coherences and of mutual information is considered as a study for inference concerning statistical graphical models. The focus of this article is simple networks. The article includes an example from hydrology.

Keywords

Graphical model, Mississippi River flow, mutual information, network, partial coherence.

1 Introduction

Science concerns relationships. The question that usually arises is what is the form of some relationship. A lesser question is how strong is a relationship. The work presented considers the use of partial coherency, and of coefficients of mutual information as measures of the strength of association of connections.

An example involving river flows measured at a succession of dams along the Mississippi River is presented. Here the nodes are in series and the edges are directed. The locations of the dams are provided in Figure 1.

Basic books discussing statistical graphical models include Cox and Wermuth [7],Whittaker [22], Edwards [9], Lauritzen [18]. The paper has the following sections: Mutual Information, Networks, Results, Discussion and Extensions.

*The research was supported by the NSF grants DMS-9704739 and DMS-9971309.

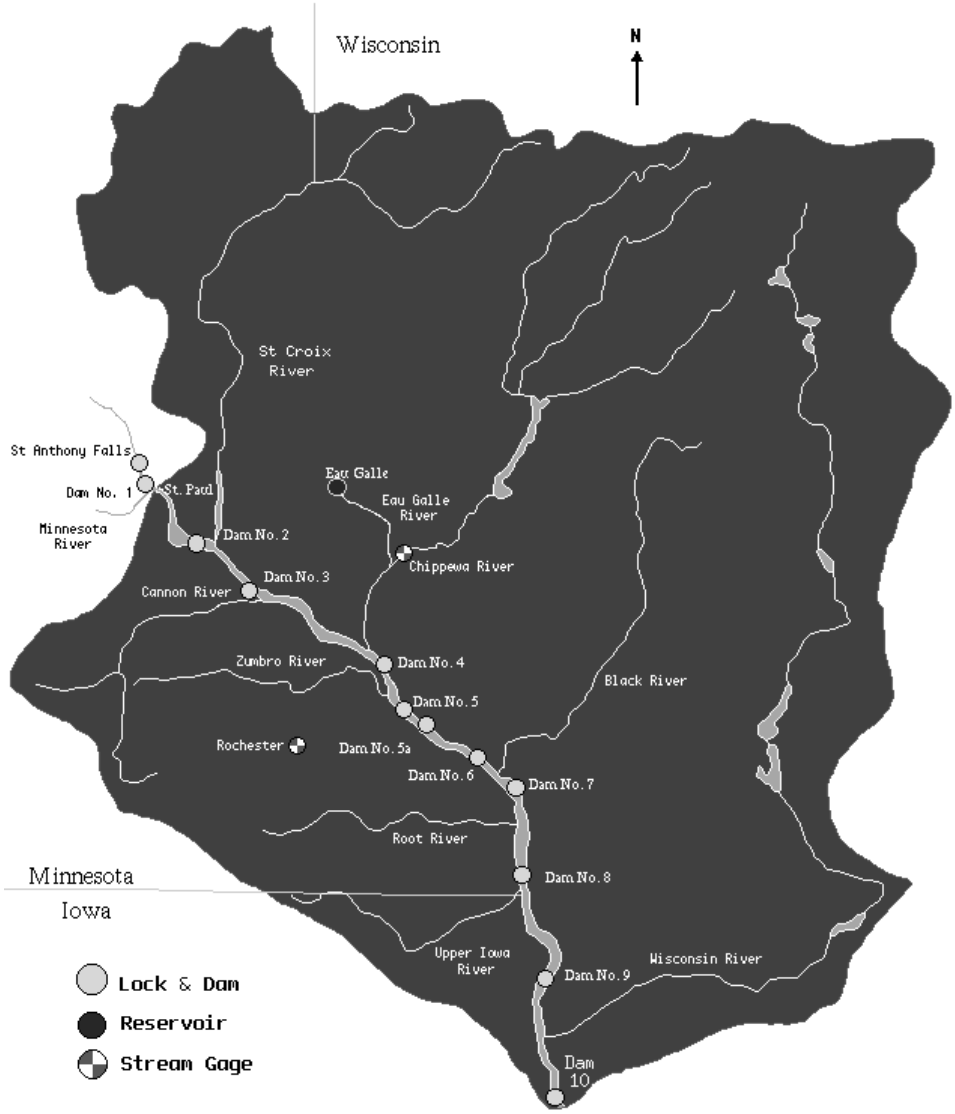


Figure 1: The locations of the 10 dams along the Mississippi River some of whose flow rates are studied.

2 Mutual Information

2.1 Continuous Case

The field of information theory provides some concepts of broad use in statistics. One of these is mutual information. It is a generalization of the coefficient of determination, $corr\{X,Y\}^2$, and it unifies a variety of problems.

For a bivariate random variable (X,Y) with density function $p(x,y)$ the mutual information (MI) is defined as

$$I_{XY} \;=\; \int_S p(x,y)\; log\frac{p(x,y)}{p_X(x)p_Y(y)}\; dxdy \tag{1}$$

where S is the region $p(x,y) \;>\; 0$.

As an example, for the bivariate normal the MI is given by

$$I_{XY} \;=\; -\frac{1}{2}log(1-\rho_{XY}^2)$$

where ρ_{XY} is $corr\{X,Y\}$.

The coefficient I_{XY} has the properties of:

1). Invariance, $I_{XY} \;=\; I_{UV}$ if the transformation $(X,Y) \rightarrow (U,V)$ has the form $U = f(X), V = g(Y)$ with f and g each 1-1 transforms.

2). Non negativity, $I_{XY} \;\geq\; 0$.

3). Measuring independence in the sense that $I_{XY} \;=\; 0$ if and only if X and Y are statistically independent.

4). Providing a measure of the strength of dependence in the senses that i) $I_{XY} \;=\; \infty$ if $Y \;=\; g(X)$, and ii) $I_{XZ} \;\leq\; I_{XY}$ if X is independent of Z given Y.

The property 3) that $I_{XY} = 0$ only if X and Y are independent stands in strong contrast to the much weaker correlation property of ρ_{XY}^2.

The estimation of entropy

There are several methods that have been used.

Nonparametric estimate

Suppose one is considering the bivariate random variable (X,Y). Supposing further that $\hat{p}(x,y)$, is an estimate of the density $p(x,y)$, for example a kernel estimate, then a direct estimate of the entropy is

$$\delta^2 \sum_{i,j}\; \hat{p}(i\delta, j\delta)\; log\; \hat{p}(i\delta, j\delta) \;\hat{=}\; E\{log\; p(X,Y)\} \quad for\; \delta\; small$$

In the same way $E\{log\; p_X(X)\},\; E\{log\; p_Y(Y)\}$ may be estimated and one can proceed to an estimate of the mutual information via expression (1). References to the type of entropy estimate just described and some statistical properties include: Joe [16], Hall and Morton [14], Fernandes [10], Hong and White, [15], Granger et al [13].

Difficulties with this form of estimate can arise when $p_X(.),\; p_Y(.)$ are small. The nonparametric form also runs into difficulty when one moves to higher dimensions.

A sieve type of estimate is presently being investigated for this situation, in particular an orthogonal function expansion employing shrunken coefficient estimates.

Parametric estimates of entropy

If the density $p(x, y|\theta)$ depends on a parameter θ that may be estimated reasonably then an immediate estimate of the entropy is provided by

$$\int p(x, y|\hat{\theta})\ log\ p(x, y|\hat{\theta})\ dxdy$$

Another form of estimate is based on the likelihood function. Suppose one has a model for the random variable (X, Y) including the parameter θ, (of dimension ν). Suppose the model has the property that X and Y are independent when $\theta = 0$. When there are n independent observations the log likelihood ratio for the hypothesis $\theta = 0$ is

$$\sum_1^n log \frac{p(x_i, y_i|\theta)}{p_X(x_i)p_Y(y_i)}$$

with expected value

$$nI_{XY}$$

This suggests the use of the loglikelihood ratio statistic divided by n as an estimate of I_{XY}. A further aspect of the use of this statistic is that its distribution will be approximately proportional to χ^2_ν, where ν is the dimension of θ, when X and Y are independent.

Partial analysis.

When networks are being considered the conditional mutual information is also of use. For a trivariate ranom variable X, Y, Z one can consider

$$I_{XY|Z} =$$

$$\int\int\int p(x, y, z)\ log\ \frac{p(x, y, z)p(z)}{p(x, z)p(y, z)}\ dxdydz$$

Its value for the trivariate normal is

$$-\frac{1}{2} log(1 - \rho^2_{XY|Z})$$

with $\rho_{XY|Z}$ the partial correlation of X and Y having removed the linear effects of Z.

2.2 Processes

A disadvantage of MI as introduced above is that it is simply a scalar. As consideration turns to the process case, i.e. functions, it seems pertinent to seeks to decompose its value somehow.

1. *Time-side approach*

The entropy of a process is defined by a suitable passage to the limit for example as

$$lim_{T\to\infty} E\{log\ p(x_1, x_2, ..., x_T)\}$$

where $p(x_1, ..., x_T)$ denotes the density of order T. To begin one can simply consider the mutual information of the values $Y(t + u), Y(t)$ or of the values $Y(t + u), X(t)$. This leads to a consideration of the coefficients

$$I_{YY}(u) \quad and \quad I_{YX}(u)$$

i.e. mutual information as a function of lag u. References to this idea include: Li [19] and Granger and Lin [12].

2. *Frequency side approach*

Similarly it seems worth considering the mutual information at frequency λ of two components of a bivariate stationary time series. This could be defined as the mutual information of $dZ_X(\lambda)$ and $dZ_Y(\lambda)$. Because these variates are complex-valued a 4-variate random variable is involved. In the Gaussian case the MI at frequency λ is

$$-\ log(1-|R_{XY}(\lambda)|^2)$$

where $|R_{XY}(\lambda)|^2$ is the coherence of X and Y at frequency λ. The overall information rate is

$$-\int_{-\pi}^{\pi} log(1-|R(\omega)|^2)d\omega$$

Granger and Hatanaka [11].

In the general case for each frequency one might construct an estimate, $\hat{I}_{XY}\ (\lambda)$, based on kernel estimates of the densities taking empirical FT-values near λ as the data. A difficulty that arises is that the random variables are complex-valued, i.e. the situation is 4-dimensional.

The way to estimate the MI, suggested above, is to fit a parameteric model and then to use the loglikelihood ratio test statistic for a test of independence.

A novel way, also being pursued, is to use first recurrence time estimates of entropy Ornstein and Weiss [20] and Wyner [23].

3 Networks.

In crude terms a *network* is a box (or node) and line (or edge) diagram and some of the lines may be directed. Some simple 3- and 4-node networks are shown in Figure 3. In our work a box corresponds to a random entity, to a random variable, to a time series or to a point process. In studying such models the methods of statistical graphical models provide pertinent methodology. Typically these models are based on conditional distributions. See the books by Edwards [9], Whittaker [22], Lauritzen [18], Cox and Wermuth [7].

If $A,\ B,\ C$ represent nodes a question may be as simple as: Is the structure $A\ \rightarrow\ B\ \rightarrow\ C$ appropriate or is it better described as $(A,B)\ \rightarrow\ C$? One the other hand the question may be as complicated as: What is the wiring diagram of the brain?

Figure 1 shows the locations of dams of a network along the Mississippi River. Since the bulk of the water flows south an elementary graphical model for this situation is: $Dam\ 1\ \rightarrow\ Dam\ 2\ \rightarrow\ ...\ \rightarrow\ Dam\ 10$. Of course there are other sources of water, such as entering rivers and rainfall to be taken note of. The figure and the data to be analyzed are taken from

www.mvp-wc.usace.army.mil/projects/lock_dam.html

One reference to the approach of this paper is Brillinger [3].

4 Results

4.1 Mississippi River Flow

The waters of the Mississippi River flow from Minnesota in the north of the United States to the Gulf of Mexico. Flooding along the river has long been a concern, and the U.S. Army Corps of Engineers has constructed a series of locks for flood control and as an aid to navigation. The waters flowing may be viewed as a system added to by precipitation and by flow from entering streams and runoff and reduced by evaporation, absorption and diversion.

The basic data employed in the study to be described are the daily water flow rates as recorded at a succession of dams along the river. The data are daily from 1960 on and were obtained from the WWW site:

www.mvp-wc.usace.army.mil/lock_dam.html

Figure 1, taken from that web site, shows the locations of the dams. Entering streams may be seen. Consider for example the logarithms of the flow rates, $Y_2(t), Y_4(t), Y_5(t)$ at Dams 2, 4, 5. Their locations may be seen in Figure 1. One sees the bulk of the waters passing from Dam 2 to Dam 4 and then onto Dam 5. One sees the Zumbo River entering between Dams 4 and 5 and the three other rivers entering between Dam 2 and Dam 4. The logarithms of the flow rates are taken as the basic variables.

This situation will be studied as providing a useful test bed for studying the effectiveness of the partial coherence and mutual information parameters.

Figure 2 presents the results of a partial coherence analysis focusing on Dams 2, 4, 5. The top right panel provides the estimated coherence functions of Dams 2 and 5. The horizontal line gives the approximate upper 95% null level under the null hypothesis of zero coherence. One sees high coherence at the low frequencies. The bottom right plot is the estimated partial coherence function of Dams 2 and 5 having removed the linear time invariant effects of Dam 4. Once again the horizontal line gives the approximate upper 95% null level under the null hypothesis of zero coherence. One looks for more than about 5% of the values lying above these lines. In the partial coherence case one sees not too much activity. In this situation the partial coherence could have been anticipated to be negligible because of Dam 4's being so close to Dam 5, i.e. highly effective in blocking off the direct effects of Dam 2. As an aid to understanding this analysis one can consider the model

$$Y_5(t) \;=\; \int a(t-u)Y_4(u)du \;+\; \epsilon(t)$$

$$Y_4(t) \;=\; \int b(t-u)Y_2(u)du \;+\; \eta(t)$$

with ϵ and η noise processes. The partial coherence estimated is then the coherence of the processes ϵ and η.

The analysis is now extended to 4 series. Figure 3 provides the estimated partial coherences of Dams 3 and 5 having removed the effects of Dam 4 and of Dams 2 and 4 having removed the effects of Dam 3 and also that of Dams 2 and 5 having removed the effects of Dams 3 and 4. The required formulas may be found in Brillinger [2]. The networks contemplated in the analysis are the parallel and series ones. Logically the series structure is appropriate.

All three of the partial coherences appear weak. This is consistent with the series structure of the graph as was anticipated. Had there been parallel links through

Dams 3 and 4 instead of the serial ones then the partial coherences 35|4 and 24|3 would not be expected to be near 0 generally.

As a further example the MI of series 2 and 5 was estimated as a function of frequency λ, actually the MI's of the two real parts of the frequency components and of the two imaginary parts. These estimates are graphed in Figure 4. Approximate 2.326 s.e. limits are obtained by randomly altering the phases of the empirical FT components, following the procedure of Braun and Kulperger [1]. In the plots one notes some apparent association at the lower frequencies. This could arise, for example, from the occurrence of snow or rain storms affecting the segments of the river at the same time.

An estimate of the full MI is currently being developed following the discussion of Section 2 .

A point process analysis of this situation is developed in Brillinger [4]. These data are also considered in Brillinger [5].

5 Discussion and extensions.

The coefficient of mutual information is a unifying concept extending second-order quantities that have restricted applicability. Its being 0 actually implies that the quantities involved are statistically independent. Another important advantage is that the MI pays no real attention to the values of the process. They can be non-negative, integers or proportions for example.

The MI is useful when one wishes to make inferences stronger than: "The hypothesis of independence is rejected." and more of the character "The strength of connection is I."

During the work the plot of the *function* $\hat{I}_{XY}(\lambda)$, appear more useful than simple scalars $\hat{I}_{XY}$. Both parametric model-based estimates and nonparametric estimates of mutual information have been mentioned and computed.

A number of extensions are available and some work is in progress. One can consider the cases of: point processes, spatial-temporal data, local estimates, of learning, of change, of trend, and of comparative experiments. Brillinger [6] contains related ideas and examples from neurophysiology.

One needs to develop the statistical properties of other estimates of MI such as the estimate based on the waiting time and the sieve estimates.

Acknowledgements

Dr. Partha Mitra made some stimulating remarks concerning the use of mutual information. The Referees remarks led to clarifications and reductions. I thank them.

Part of the material was presented as the Opening Lecture at the XXXIème Journées de Statistique in Grenoble in 1999.

References

[1] Braun, J. and Kulperger, R. (1997). Properties of a Fourier bootstrap method for time series. *Commun. Statist.-Theory Meth.* 26, 1329-1336.

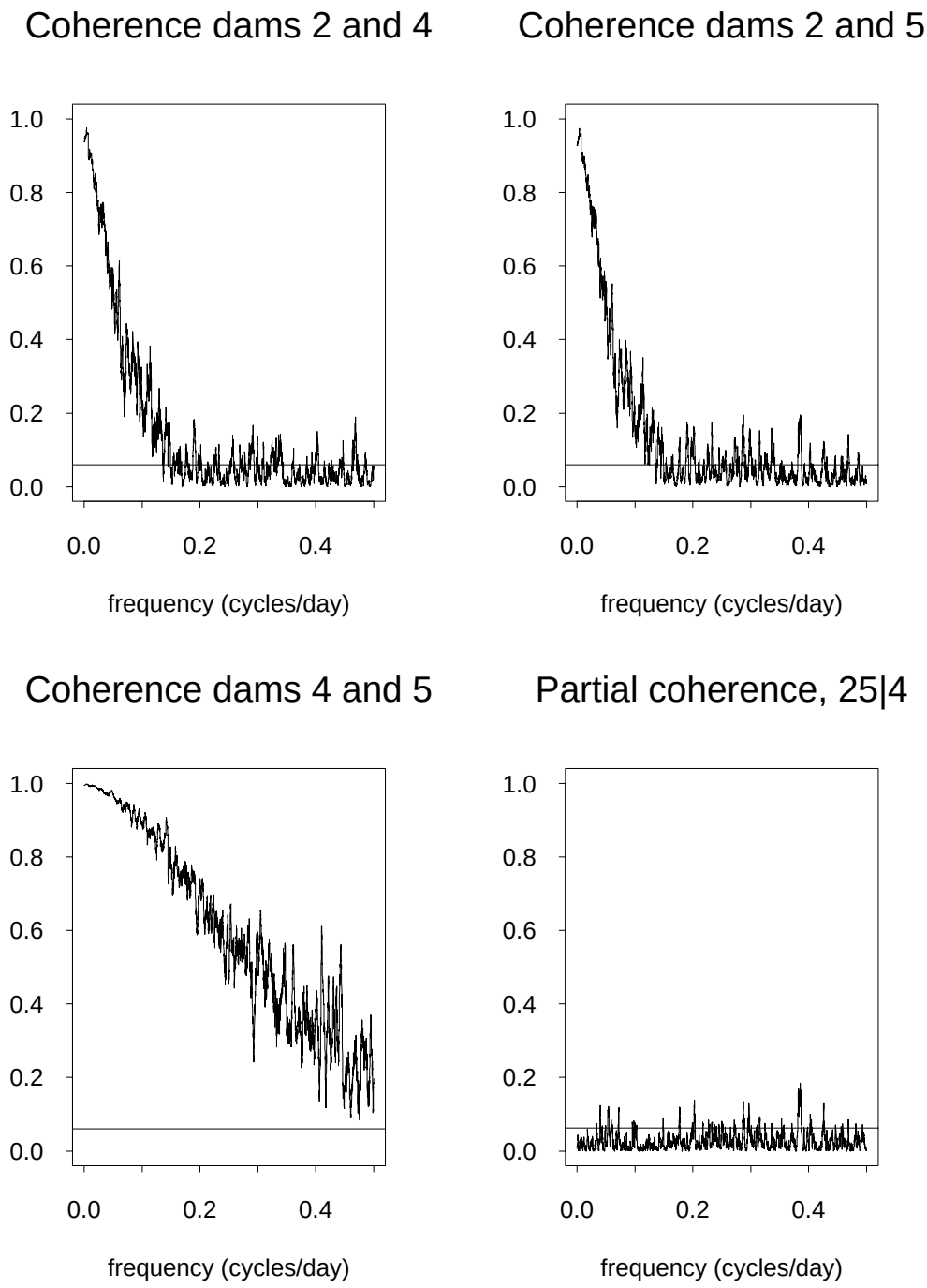


Figure 2: Estimated partial coherence of log(flow rate) at Dams 2 and 5 given those at Dam 4. The horizontal line gives the approximate upper 95% null level.

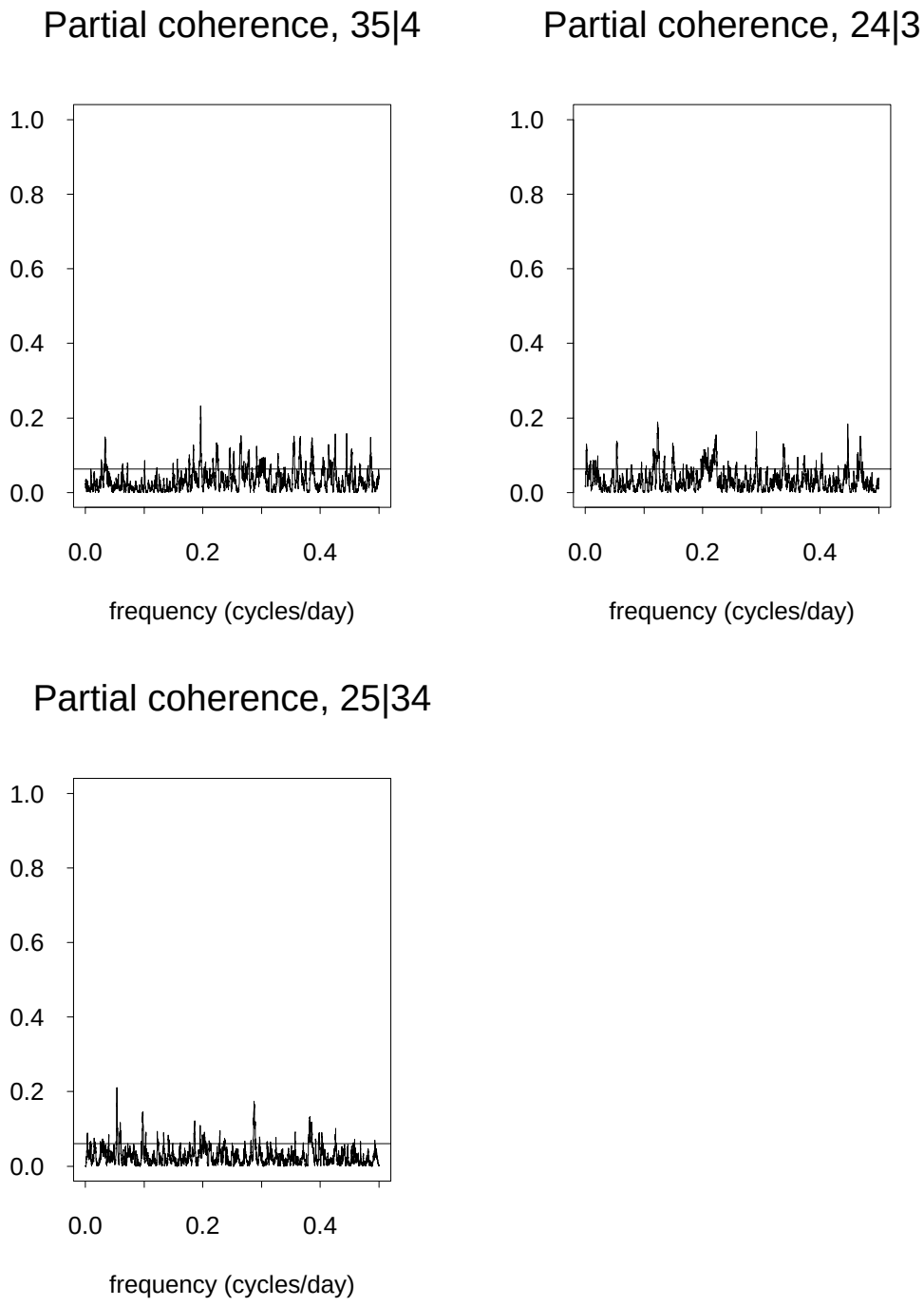
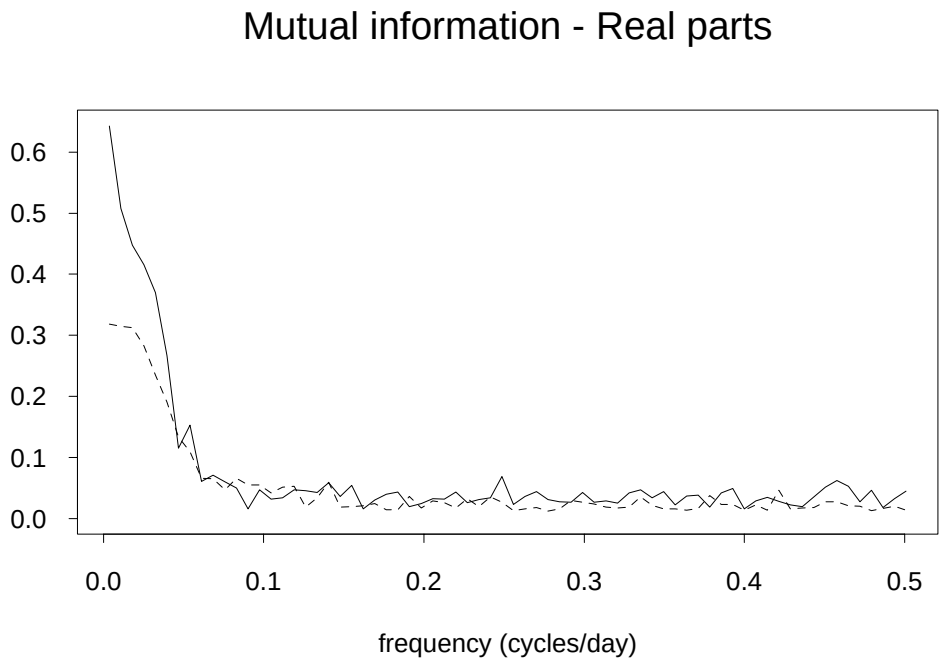


Figure 3: The estimated partial coherence of log(flow rate) at Dams 3 and 5 given Dam 4, Dams 2 and 4 given Dam 3 and Dams 2 and 5 given those at Dams 3 and 4. The horizontal lines give the upper 95% null level.



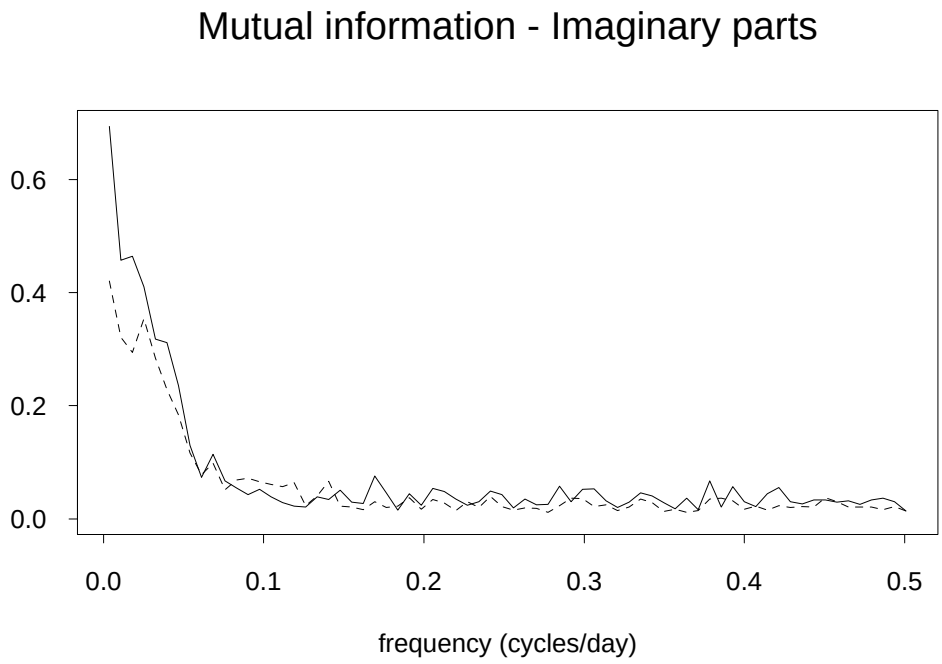


Figure 4: The top panel provides an estimate of the MI of the two real parts of the Mississippi river flows at Dams 2 and 5. The bottom panel similarly refers to the two imaginary parts. The dashed line gives the upper 95% null level.

[2] Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory.* Holt-Rinehart, New York. Republished as a SIAM Classic in Applied Mathematics (2001).

[3] Brillinger, D. R. (1996). Remarks concerning graphical models for time series and point processes. *Brazilian Review Econometrics* 16, 1-23.

[4] Brillinger, D. R. (2001). Does anyone know when the correlation coefficient is useful? A study of the times of extreme river flows. *Technometrics* 43, 266-273.

[5] Brillinger, D. R. (2001). John Tukey and the correlation coefficient. *Proc. Interface 2001.*

[6] Brillinger, D. R. (2002). Modelling and analysis of some random process data from neurophysiology. To appear in *Revista Investigacion Operacional.*

[7] Cox, D. R. and Wermuth, N. (1998). *Multivariate Dependencies: Models, Analysis, and Interpretation.* Chapman & Hall, London.

[8] Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* Wiley, New York.

[9] Edwards, D. (1995). *Introduction to Graphical Modelling.* Springer, New York.

[10] Fernandes, M. (2000). Nonparametric entropy-based tests of independence between stochastic processes. Preprint, Fundacão Getulio Vargas, Rio de Janeiro.

[11] Granger, C. W. J. and Hatanaka, M. (1964). *Spectral Analysis of Economic Time Series.* Princeton University Press, Princeton.

[12] Granger, C. W. J. and Lin, J-L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Series Anal.* 15, 371-384.

[13] Granger, C. W. J., Maasouni, E. and Racine, J. (2000). A dependence metric for nonlinear time series. Preprint. Dept. of Economics, UCSD.

[14] Hall, P. and Morton, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* 45, 69-88.

[15] Hong, Y. and White, H. (2000). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. Preprint, Economics Department, UCSD.

[16] Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* 41, 683-697.

[17] Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics* Vol. 2. Griffin, London.

[18] Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press, Oxford.

[19] Li, W. (1990). Mutual information functions versus correlation functions. *J. Statistical Physics* 60, 823-837.

[20] Ornstein, D. S. and Weiss, B. (1993). Entropy and data compression schemes. *IEEE Inf. Theory* 39, 78-83.

[21] Rao, C. R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, New York.

[22] Whittaker, (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.

[23] Wyner, A. J. (1999). More on recurrence and waiting times. *Ann. App. Prob.* 9, 780-796.