ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Mutual information in the frequency domain

David R. Brillinger*, Apratim Guha

Statistics Department, University of California, Berkeley, CA 94720-3860, USA

Abstract

Coefficients of mutual information (MI) can provide powerful extensions of classical coefficients of correlation. In particular, they have the property of vanishing if and only if the components involved are statistically independent of each other. This characteristic can prove useful in preparatory work to model building. In this article a frequency domain variant of MI is developed and studied for bivariate stationary time series. As a scientific example an ambient seismic noise data set is studied and a lack of independence of the components inferred. The character of the dependence of the MI on frequency may be used to suggest the nature of the statistical dependence.

MSC: Primary, 62M15;; 94A17; secondary 62M99; 94A12

Keywords: Coherence; Correlation; Frequency domain; Mutual information; Time series

1. Introduction

Frequency domain analysis have proved successful in a wide variety of situations. Consider a bivariate stationary time series, $(X(t), Y(t))$, and its frequency analysis. When the second moments exist the series has a Cramér representation

$$X(t) = \int_{-\pi}^{\pi} \exp\{\mathrm{i}\lambda t\}\,\mathrm{d}Z_X(\lambda), \quad Y(t) = \int_{-\pi}^{\pi} \exp\{\mathrm{i}\lambda t\}\,\mathrm{d}Z_Y(\lambda)$$

for $t = 0, \pm 1, \pm 2, \ldots$, where the bivariate process (Z_X, Z_Y) has orthogonal increments and contains the statistical properties of the series itself, (Brillinger, 1975). The integral notation appearing is not a classical Riemann–Stieltjes one, rather it is a symbolic representation of a variate existing as a limit in mean. The reason for the notation is that the limit possesses some of the properties of an ordinary integral, e.g. additivity. The components, Z_X and Z_Y, are complex-valued.

The *coherency* at frequency λ of the series is defined as

$$R_{XY}(\lambda) = \mathrm{corr}\{\mathrm{d}Z_X(\lambda), \mathrm{d}Z_Y(\lambda)\},$$

where for the complex-valued variate (U, V) the quantity $\mathrm{corr}\{U, V\}$ is taken to be

$$E\{[U - E\{U\}]\overline{[V - E\{V\}]}\}/\sqrt{E\{|U - E\{U\}|^2\}E\{|V - E\{V\}|^2\}}.$$

* Corresponding author. Tel.: +1 510 642 0611; fax: +1 510 642 7892.
E-mail address: brill@stat.berkeley.edu (D.R. Brillinger).

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.06.026

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

The *coherence* at frequency λ is the modulus-squared

$$|R_{XY}(\lambda)|^2.$$

It is a measure of the degree of linear time invariant association of the pair of series $\{X(t)\}$ and $\{Y(t)\}$. One motivation for the use of coherence concerns relationships of the form

$$Y(t) \approx \mu + \sum_{u=-\infty}^{\infty} a(u)X(t-u), \quad t = 0, \pm 1, \pm 2, \ldots \tag{1}$$

for constant μ and $\{a(u)\}$. Coherence values lie in the interval [0,1] with the value 0 occurring when the two series are uncorrelated at all lags and the value 1 occurring when equality occurs in (1). There is further discussion of the coherence parameter and linear time invariant relations (filters) in Brillinger (1975).

However, the coherence is inadequate as a measure of general association for it may be identically 0 when two series are in fact related. This parallels the circumstance for the ordinary coefficient of correlation which is itself a measure of the strength of linear association. An example of 0 coherence, yet related series, is the following. Let ζ represent a real-valued random variable with $E\{\zeta\}$, $E\{\zeta^3\} = 0$. Consider the two series

$$\delta(t) = \sum_u a(t-u)\zeta(u), \quad \varepsilon(t) = \sum_u b(t-u)\zeta(u)^2 \tag{2}$$

for the $\zeta(u)$ independent realizations of the variate ζ and for filters $\{a(u)\}$, $\{b(u)\}$ such that the series $\{\delta(t)\}$ and $\{\varepsilon(t)\}$ are well-defined. The coherence of the series δ and ε is identically 0, yet they are statistically related through common dependence on $\{\zeta(t)\}$.

Such behavior does not occur for the coefficient of mutual information (MI). This coefficient has the property of taking on the value 0 if and only if the variables are statistically independent.

The article proceeds to a study of MI in the frequency domain by working with the spectral increments $\mathrm{d}Z_X(\lambda)$ and $\mathrm{d}Z_Y(\lambda)$. Two-, three- and four-dimensional parameters and statistics are considered.

In a preliminary empirical study the MI as a function of frequency is estimated for a lengthy, bivariate series, specifically seismic noise recorded at the Farallon Islands off San Francisco.

Madan Puri has been enriching the literature of statistics during our careers. He has worked on many aspects of statistics including both information measures and time series analysis. The Current Index of Statistics lists some 11 of his papers on time series analysis during the period 1984–1995. It is an honor to be invited to contribute to this volume.

The sections of the paper are: Some review, Estimation in the frequency domain, The data, Results, Discussion and summary.

2. Some review

MI is a tool for the study of statistical dependence. It was first introduced by Shannon (1948), and there have been substantial developments and applications of the concept since. One may mention Cover and Thomas (1991), Kantz and Schreiber (1997), Pluim et al. (2003) and references therein.

When the partitioned random variable $(\mathbf{U}, \mathbf{V})$ has a continuous distribution, the *mutual information* is defined as

$$I_{\mathbf{U},\mathbf{V}} = \int\int \log\left(\frac{p_{\mathbf{UV}}(\mathbf{u},\mathbf{v})}{p_{\mathbf{U}}(\mathbf{u})p_{\mathbf{V}}(\mathbf{v})}\right) p_{\mathbf{UV}}(\mathbf{u},\mathbf{v})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}. \tag{3}$$

A popular MI estimator in this case is the "plug-in" estimator (Antos and Kontoyiannis, 2001), or "naive" estimator (Strong et al., 1998). This estimate is obtained by substituting suitable density estimators into (3). In the bivariate case it takes the form

$$\hat{I}_{U,V} = \int\int \log\left(\frac{\hat{p}_{UV}(u,v)}{\hat{p}_U(u)\hat{p}_V(v)}\right) \hat{p}_{UV}(u,v)\,\mathrm{d}u\,\mathrm{d}v. \tag{4}$$

The density estimators could be parametric (Brillinger, 2004) or nonparametric. The nonparametric density estimators may be based either on histograms (Moddemeijer, 1989) or on continuous kernels (Mars and van Aragon, 1982;

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

Joe, 1989; Granger and Lin, 1994; Moon et al., 1995). Typically the estimates are biased. Darbellay (1998) develops an estimate in which the cells may have different widths and in which parameters of the estimation procedure are determined in automatic fashion.

Consider as well the discrete case, often used as an approximation. Take a bivariate discrete chance quantity (U, V) with U taking on the values $1, \ldots, J$ and V the values $1, \ldots, K$ and define

$$\text{Prob}\{U = j, V = k\} = p_{jk}.$$

Writing the marginals as p_{j+}, p_{+k}, the MI in this case is

$$I_{UV} = \sum_{j,k} p_{jk} \log \frac{p_{jk}}{p_{j+}p_{+k}}, \tag{5}$$

assuming that $p_{jk} \neq 0$. Turning to the problem of estimation, represent the variate (U, V) by $W = \{W_{jk}\}$ with $W_{jk} = 1$ if the result (j, k) occurs and $W_{jk} = 0$ otherwise. Suppose that there are n independent realizations, $\{w_{jkl}, l = 1, \ldots, n\}$, of W. The maximum likelihood estimates of the p_{jk} are the $\hat{p}_{jk} = \sum_l w_{jkl}/n$ and the plug-in estimate of the MI is

$$\hat{I}_{UV} = \sum_{j,k} \hat{p}_{jk} \log \frac{\hat{p}_{jk}}{\hat{p}_{j+}\hat{p}_{+k}}. \tag{6}$$

The histogram estimate has this form. The quantity (6) is $1/2n$ times the G^2 statistics employed by researchers in contingency tables. The null distribution of G^2 is approximately $\chi^2_{(J-1)(K-1)}$, see Christensen (1997) for example.

With cells of width $\Delta u \Delta v$ and $\hat{p}_{jk}$ the proportion of observations in cell (j, k), the quantity on the right of (6) would be multiplied by $\Delta u \Delta v$ to provide an estimate of (3).

Mutual information has been employed to assess the dependence between the components, $\{X(t)\}$ and $\{Y(t)\}$, of a stationary time series, by looking at the MI of $X(t + u)$ and $Y(t)$ as a function of lag u. Moddemeijer (1989) provides expressions of asymptotic bias and variance for a histogram estimator such as (6) in this case. He assumes that the pairs $(X(t + u), Y(t))$ are related amongst themselves but $X(t + v)$ is independent of $Y(t)$ for $v \neq u$. Moddemeijer (1999) provides an expression of the variance when the samples are not independent.

Among the major contributors to the study of MI estimates, Antos and Kontoyiannis (2001) proved the consistency of the nonparametric plug-in estimator (5) based on a histogram-type estimate in the discrete setup. They showed that such an estimator is $\sqrt{n}$-consistent (i.e. rate of convergence of the estimator to the original value is proportional to the square root of sample size) when the population is finite. In general, this rate is not achieved for an infinite population. They also obtained some rates for infinite populations under various extra regularity conditions. Aspects by which the various estimates may be compared include ease of choice of parameters appearing in the definition and computational and statistical complexity.

For the parametric case, Brillinger (2004) proves that, under regularity conditions, the statistic $I_{UV}(\hat{\theta})$ is $\sqrt{n}$-consistent and asymptotically normal when U and V are dependent and $\partial I_{U,V}(\theta)/\partial\theta$ is non-zero at $\theta = \theta_0$, the true value. Here $\hat{\theta}$ is the maximum likelihood estimate. Other statistics are also discussed there and shown to have asymptotic χ^2 distributions under independence of the two variables.

Joe (1989) shows that for bounded continuous U and V, and further regularity conditions, the estimator (4) is consistent. Among other things, Joe finds an expression for the mean square error.

3. Estimation in the frequency domain

Consider the bivariate stationary time series $(X(t), Y(t))$ and suppose that one is interested in the strength of association between $X(t)$ and $Y(t)$ at frequency λ.

In the case of a bivariate stationary Gaussian series, the MI of $X(t)$ and $Y(t)$ is

$$-\frac{1}{4\pi}\int_0^{2\pi} \log[1 - |R_{XY}(\lambda)|^2]\,\mathrm{d}\lambda$$

(Gelfand and Yaglom, 1959). This provides a connection with the coherence function defined earlier in the article.

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

The series $X(t)$ and $Y(t)$ will be statistically independent only if the components $Z_X(\lambda)$ and $Z_Y(\lambda)$ are statistically independent for all λ. This possibility may be studied by means of mutual information. In what follows three cases are considered. These are (a) the mutual information I_{UV} for the real-valued variates

$$U = |\mathrm{d}Z_X(\lambda)|^2, \quad V = |\mathrm{d}Z_Y(\lambda)|^2; \tag{7}$$

(b) $I_{\mathbf{U}V}$ for the bivariate $\mathbf{U}$ and the real-valued V

$$\mathbf{U} = (\mathrm{Re}\{\mathrm{d}Z_X(\lambda)\}, \mathrm{Im}\{\mathrm{d}Z_X(\lambda)\}), \quad V = \mathrm{Re}\{\mathrm{d}Z_Y(\lambda)\}; \tag{8}$$

(c) $I_{\mathbf{UV}}$ for the bivariate $\mathbf{U}$ and the bivariate $\mathbf{V}$

$$\mathbf{U} = (\mathrm{Re}\{\mathrm{d}Z_X(\lambda)\}, \mathrm{Im}\{\mathrm{d}Z_X(\lambda)\}), \quad \mathbf{V} = (\mathrm{Re}\{\mathrm{d}Z_Y(\lambda)\}, \mathrm{Im}\{\mathrm{d}Z_Y(\lambda)\}). \tag{9}$$

These involve, respectively, two, three and four real-valued variates.

Proceeding to the estimation of the MI for such quantities, one can use the empirical Fourier transform (FT) of a stretch of time series to estimate the spectral increments $\mathrm{d}Z$, see Brillinger (1975, Section 4.6). There one finds the motivating approximation

$$Z_X^{(T)}(\lambda) = \frac{1}{2\pi} \sum_{t=-T}^{T} X(t)[1 - \exp\{-\mathrm{i}\lambda t\}]/(-\mathrm{i}t).$$

The estimates computed below use such FTs splitting the data into contiguous time segments of 4000 points, computing the Fourier transforms of the individual stretches and then employing the histogram form (6) with 10 cells. Ten cells were employed in this preliminary study for computational convenience.

In the case of the variates (7) the estimate is in essence based on the periodograms of the segments.

Brillinger (2002) estimates the MI of $\mathrm{Re}\{\mathrm{d}Z_X(\lambda)\}$ with $\mathrm{Re}\{\mathrm{d}Z_Y(\lambda)\}$, and of the corresponding imaginary parts, for Mississippi river flow at two dams. In each case one notes significant values of the MI estimate up to a frequency of about .05 cycles/day. It was remarked there that this association might arise from snow or rain storms affecting the regions of the two dams at the same time.

4. The data

The frequency domain parameters just referred to are estimated for a stretch of seismic noise. Seismic noise data refer to time series recorded by a seismometer when there are, apparently, no earthquakes taking place. In many cases the noise is thought to be due to ocean waves. There is some discussion of ambient seismic noise in Aki and Richards (1980, p. 497–498), They present a "representative" estimated power spectrum and refer to peaks around .07 and .14 Hz (cycles per second), respectively.

The data studied were measurements obtained during a 12 h period on September 2, 2004, by an instrument on the Farallon Islands off San Francisco. These particular data are studied because they are from an island, because there are possibilities of non-Gaussian and nonlinear behavior, and because lengthy time series may be collected. Velocity of the ground motion was recorded as a function of time in three perpendicular directions. In the work presented the two horizontal components are employed. A lowpass filter was used, with cutoff at 16 Hz. This was to reduce the effects of aliasing, see Brillinger (1975). The spacing between the digital measurements was $\frac{1}{40}$ s, leading to $T = 1,728,000$ data points in all.

In the analysis presented the series X is the horizontal east–west velocity and Y the north–south. The data were first graphed in order to assess stationarity and to scan for outliers. In this stage the series were divided into contiguous segments of 1000 points and the median value of each segment computed. Fig. 1 provides the plots of the time series segments $X(t)$ and $Y(t)$, $t = 1, \ldots, T$. The series evidence no substantial departure from stationarity.

5. Results

To reduce leakage in frequency domain analysis series are tapered before computing the Fourier transform. Here the segments contained 4000 observations. Power spectra were estimated up to the Nyquist frequency of 20 Hz. The

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

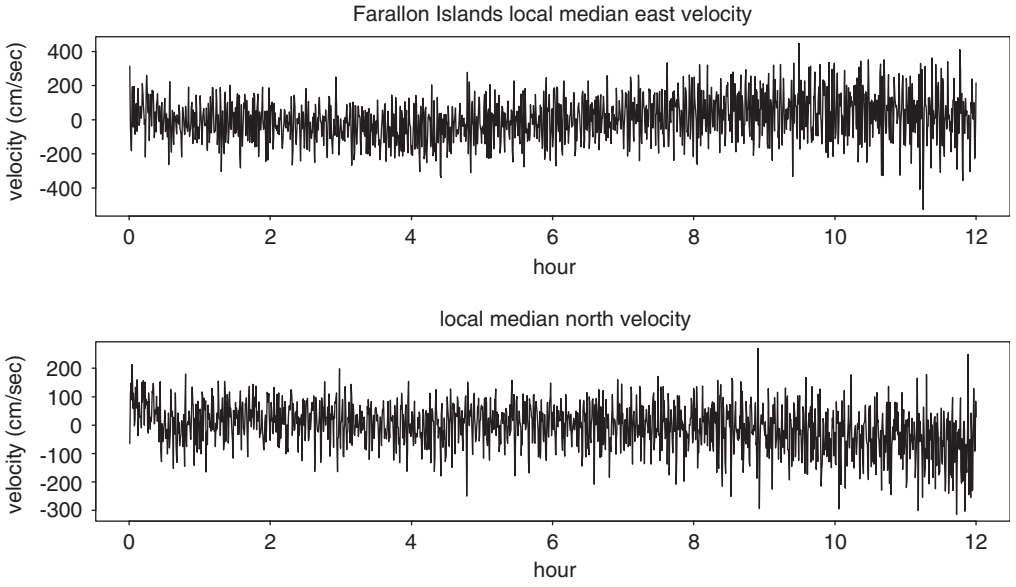

Fig. 1. The medians of contiguous segments of 1000 points of the east–west and north–south components of the ground velocity recorded at the Farallon Islands on September 2, 2004.

estimates showed a sharp cutoff near 16 Hz as was to be anticipated from the anti-aliasing filtering the seismologist employed. They further showed a substantial dropoff in power past the frequency of .3 Hz, and for this reason the figures that follow consider only the frequencies in the band from 0 to 1.0 Hz. The top half of Fig. 2 provides the estimated power spectra of the two components.

The estimated east and north spectra are nearly identical, in particular in displaying three peaks. The lower two peaks may be thought of as the .07 and .14 Hz ocean wave phenomena mentioned in Aki and Richards (1980). The source of the peak around .3 Hz requires further consideration.

The lower left panel of the figure is the coherence estimate. The horizontal line in the panel is the approximate upper 95% point of the null distribution of the coherence statistic, see Brillinger (1975, Exercise 8.6.22). The graph suggests the presence of some linear time invariant association between the two components at frequencies below 0.4 Hz. This might be the result of a signal arriving that is common to the east and north components.

The lower right panel provides the estimated phase. The phase is useful for studying leads or lags between two stationary series, but its estimate is highly variable when the coherence is low, as it is in the case at hand.

The inference then is that the series are associated, in a linear time invariant way, at frequencies below .4 Hz.

Consideration now turns to inferences based on estimates of mutual information. It is to be remembered that, in contrast to the coherence function, the MI is 0 if and only if the component variates are statistically independent. Fig. 3 presents the estimated mutual information for the variates of (7). The estimate is based on periodograms of corresponding segments of the two series. The number of X and Y cells employed is 10 here and in the computations below. The estimated MI is the curve rising above the collection of noise-like curves at the base of the figure.

The noise-like curves are meant to suggest the sampling variability in the null case of independence. To compute them the segments of the X series are randomly permuted and then the MI estimate between the resulting series and Y computed. This procedure has the advantage that the power spectral estimates remain the same as the original ones. This was done 19 times in order to obtain a null distribution test procedure of size $95\% = 100 * \frac{19}{20}$. The result and its apparent significance came as a surprise, specifically the apparently high MI in the frequency band of .2–.4 Hz. It was surprising because the coherence plot of Fig. 2 did not suggest strong association in that band. An inference might be that the frequencies from .2 to .4 Hz are associated in some strictly nonlinear sense.

The MI estimate did not change much when the logarithms of the periodograms were employed. The logarithm was considered because of its variance stabilization property for the periodogram.

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

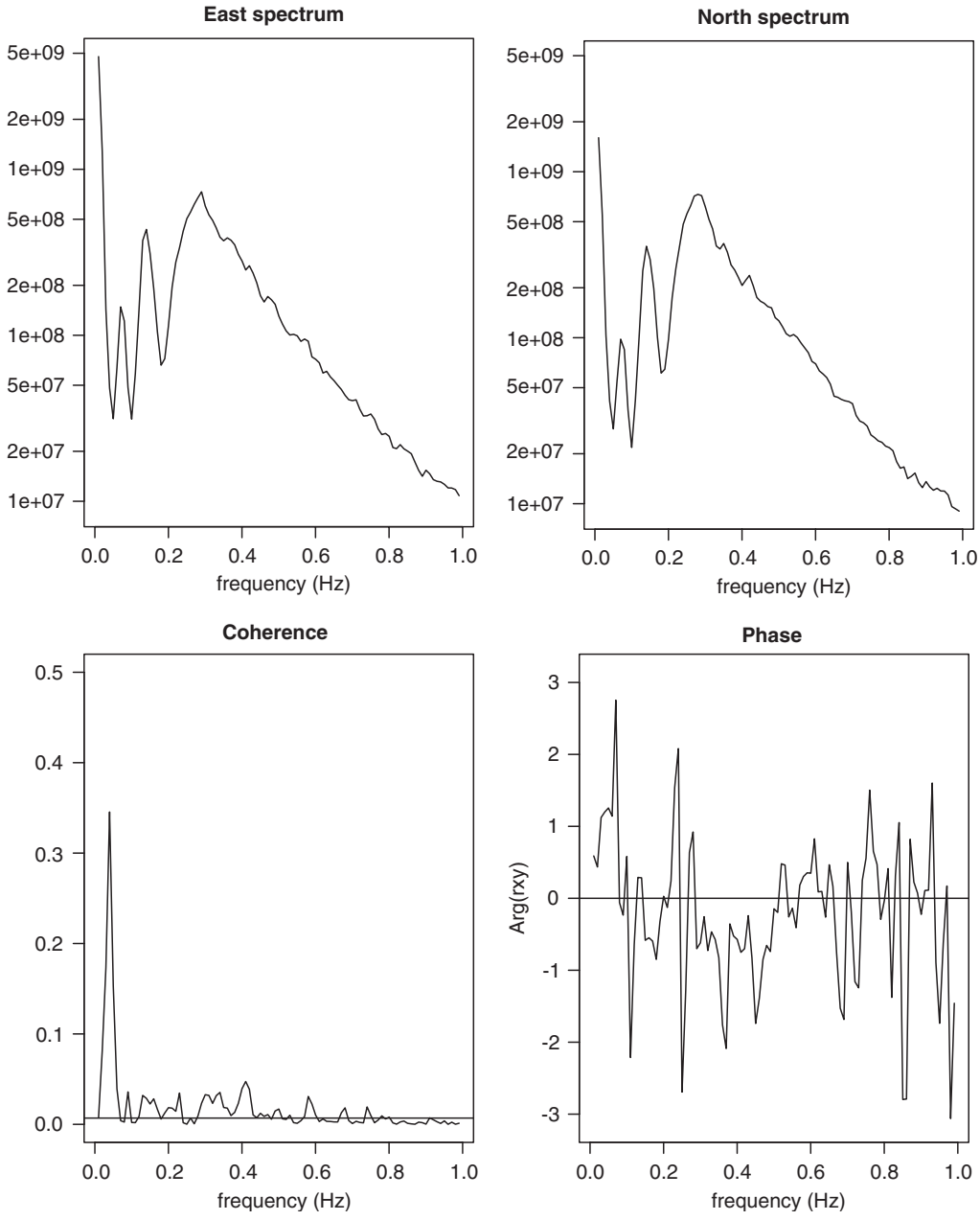

Fig. 2. The results of a spectrum analysis of the east and north series. The series were segmented into stretches of length 4000, the segments tapered, and the FTs computed. The periodograms and the crossperiodogram were averaged to estimate the second-order spectra. The horizontal line in the lower left panel is an approximate upper 95% null level.

Fig. 4 shows the estimated MI of the variates of (8) again with 19 null permutation runs superposed. There was another surprise, the similarity of Fig. 3 to Fig. 4. The variability of the MI estimate, at least in the null case, appears approximately proportional to the estimate. If an approximation of the null distribution based on a $\chi^2_{(JK-1)(L-1)}$ had been employed, it would have been constant across all frequencies. It might have been anticipated that the variability of the estimate would depend on the autocorrelation properties of the component series.

Fig. 5 shows the estimate MI of the variates (9). The just referred to behavior continues. One is left inferring that the variates $Z_X(\lambda)$ and $Z_Y(\lambda)$ are statistically dependent for a range of frequencies.

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

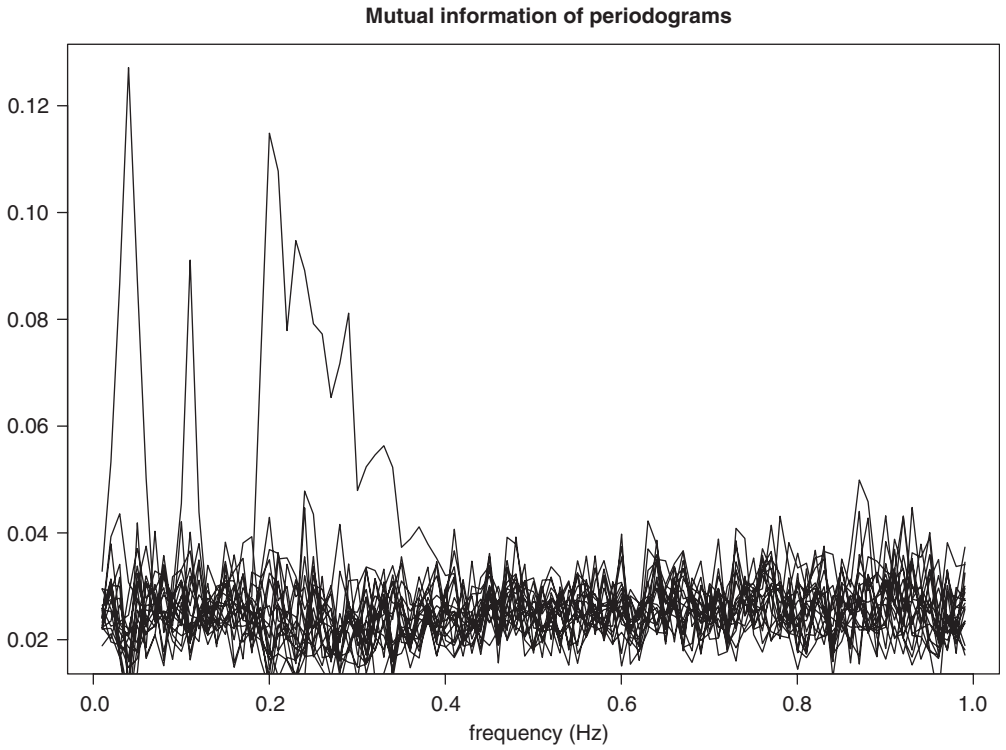

Fig. 3. The estimated MI of the variates of (7). It is based on corresponding periodograms of the component series. The lower superposed noise-like curves are the results of the 19 permutation runs referred to in the text.

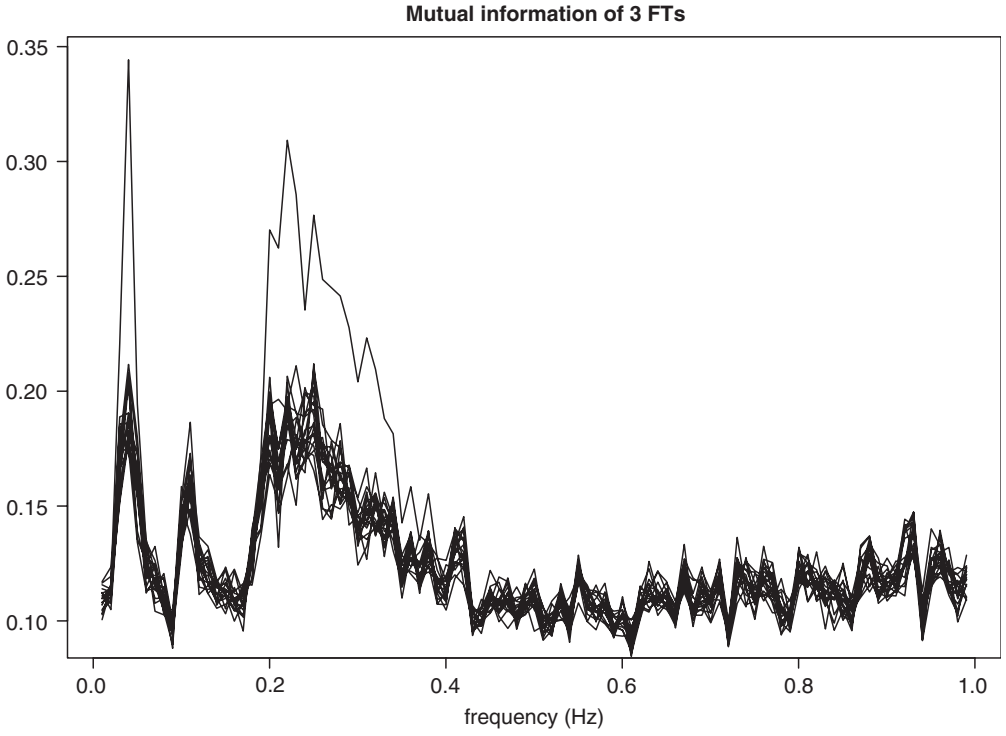

Fig. 4. The estimated mutual information of the variates of (8).

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

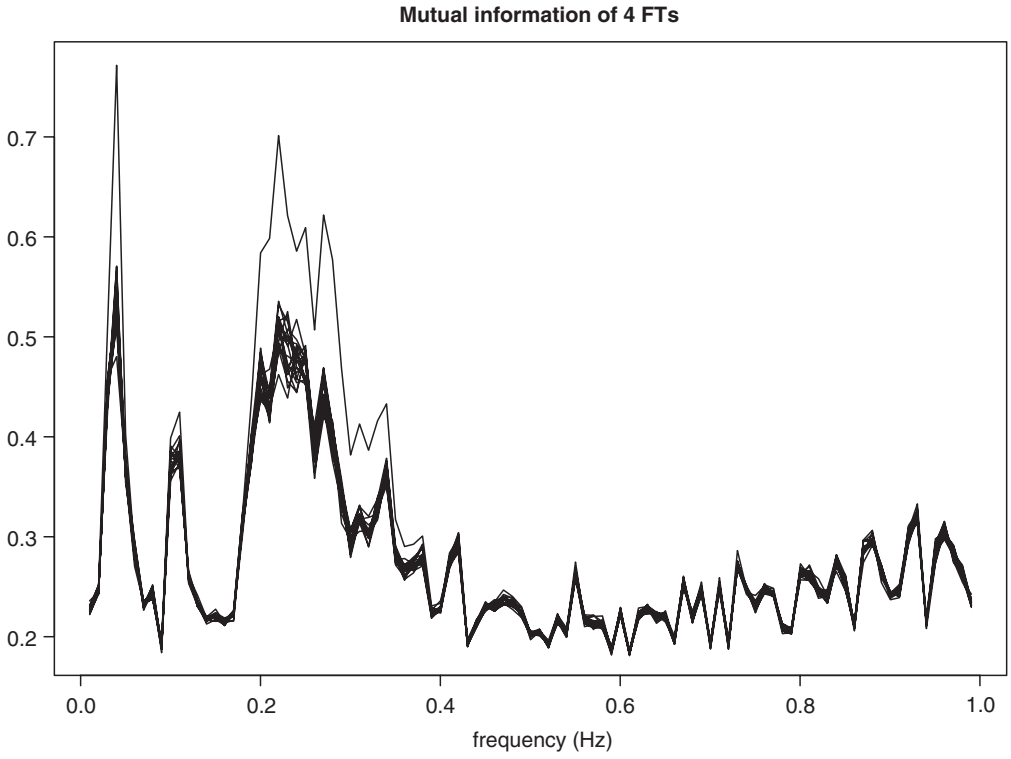

Fig. 5. The estimated mutual information of the variates of (9). Again 19 null permutation runs have been superposed.

As part of the study of the validity of the results, prewhitening by fitting long autoregressives and stronger tapers was carried out, yet the phenomena remained. The concern was that perhaps the estimates suffered from some form of spectral leakage.

6. Discussion and summary

It has been shown that mutual information analysis in the frequency domain is possible and can lead to interesting results. MI is a better quantity to employ than the coherence function of cross-spectral analysis because one can infer independence not just coherence 0.

There have been some surprises. One is the low level of association in the frequency band .1–.4 Hz became much increased in the MI plots.

The discussion of results is a bit tricky because given a time series whose support is a particular frequency band, that band can be changed by simple transformation of the series, e.g. working with $X(t)^3$ instead of $X(t)$.

The results presented are preliminary and need to be interpreted with care. Future work is needed. Practical details of the estimates need to be studied further, for example the lengths of the segments to be employed and the number of cells in the estimates of MI.

Acknowledgments

The work was supported by the NSF Grants DMS-0203921 and DMS-0504162. The authors thank Dr. Bob Uhrhammer of the Berkeley Seismographic Laboratory for providing the Farallon Islands data and the referees for helpful comments.

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026

References

Aki, K., Richards, P.G., 1980. Quantitative Seismology. Freeman, San Francisco.

Antos, A., Kontoyiannis, Y., 2001. Convergence properties of functional estimates for discrete distributions. Random Structures and Algorithms 19, 163–193.

Brillinger, D.R., 1975. Time Series: Data Analysis and Theory. Holt New York. Reprinted in 2001 as a SIAM Classic in Applied Mathematics.

Brillinger, D.R., 2002. Second-order moments and mutual information in the analysis of time series. In: Chaubey, Y.P. (Ed.), Recent Advances in Statistical Methods. Imperial College Press, London, pp. 64–76.

Brillinger, D.R., 2004. Some data analyses using mutual information. Brazilian J. Probab. Statist. 18, 163–183.

Christensen, R., 1997. Log-linear Models for Logistic Regression. Springer, New York.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Darbellay, G., 1998. An adaptive histogram estimator for mutual information. UTIA Research Report 1889. Academy of Sciences, Prague.

Gelfand, I.M., Yaglom, A.M., 1959. Calculation of the amount of information about a random function contained in another such function. Amer. Math. Soc. Transl. Ser. 2 (12), 99.

Granger, C.W.J., Lin, J.-L., 1994. Using the mutual information coefficient to identify lags in nonlinear models. J. Time Series Anal. 15, 371–384.

Joe, H., 1989. Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Statist. Math. 41, 683–697.

Kantz, H., Schreiber, T., 1997. Nonlinear Time Series Analysis. Cambridge, Cambridge.

Mars, N.J.I., van Aragon, G.W., 1982. Time delay estimation in non-linear systems using average amount of mutual information analysis. Signal Process. 4, 139–153.

Moddemeijer, R., 1989. On estimation of entropy and mutual information of continuous distributions. Signal Process. 16, 233–248.

Moddemeijer, R., 1999. A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations. Signal Process. 75, 51–63.

Moon, Y.I., Rajagopalan, B., Lall, U., 1995. Estimation of mutual information using kernel density estimators. Phys. Rev. E 52, 2318–2321.

Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A., 2003. Mutual information based registration of medical images: a survey. IEEE Trans. Med. Imag. XX, 1–21.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Tech. J. 27, 379-423, 623-656.

Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R., Bialek, W., 1998. Entropy and information in neural spike trains. Phys. Rev. Lett. 80, 197–200.

Please cite this article as: David R. Brillinger, Apratim Guha, Mutual information in the frequency domain, Journal of Statistical Planning and Inference (2006), doi: 10.1016/j.jspi.2006.06.026