Testing Nonstationary Time Series for Gaussianity and Linearity using the Evolutionary Bispectrum: An application to Internet Traffic Data

Subba Rao, T and Tsolaki, E

2006

MIMS EPrint: **2006.43**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Testing Nonstationary Time Series for Gaussianity and Linearity using the Evolutionary Bispectrum: An application to Internet Traffic Data

T. Subba Rao & Eleni P. Tsolaki

First version: 23 September 2005

Research Report No. 7, 2005, Probability and Statistics Group
School of Mathematics, The University of Manchester

Testing nonstationary time series for Gaussianity and linearity using the evolutionary bispectrum: An application to internet traffic data

T. Subba Rao

School of Mathematics,

The University of Manchester

P.O. Box 88, Manchester, M60 1QD, United Kingdom

e-mail: tata.subbarao@manchester.ac.uk

Eleni P. Tsolaki

Department of Applied Mathematics,

University of Crete, GR-714 09 Heraklion, Crete, Greece

e-mail : tsolaki@tem.uoc.gr

July 13, 2005

Abstract

We propose statistical tests for Gaussianity and linearity of nonstationary time series based on the evolutionary bispectrum. These tests can be applied to a particular subclass of nonstationary processes, the so-called oscillatory (also known as slowly varying) processes. We then apply these tests to time series of network measurements arising from internet traffic. Recent works by several researchers have demonstrated that such internet traffic processes are typically nonstationary. Also, the question of whether such processes can be described by some model whose parameters vary with time has been raised and studied at some length. We use the tests developed in this paper to show that there is evidence of non Gaussianity and nonlinearity in such processes under the assumption that they are described by a model whose parameters (and so its spectral characteristics) vary slowly with time.

1 Introduction

Spectral methods are often used in the analysis of a time series. Under the assumption that a given time series is Gaussian, second order spectral meth-

ods are sufficient. Many time series however are not necessarily Gaussian (or even linear) and as a result higher order spectral (HOS) methods are required in analysing them. A detailed systematic study of higher order spectra (cumulant spectra) can be found in [2], [3], [18] and applications of these studies have been reported in [8] and [21]. Applications of HOS to digital signal processes have been given in [11], [26] and [27].

It is to be emphasized that all of the above methods depend heavily on the assumption that the series are stationary and that the linear systems considered are time invariant. It is also to be expected that in practice, the assumption of stationarity may sometimes be unrealistic. In order to overcome these problems, Priestley [12] introduced a spectral representation for a class of nonstationary processes, the so-called oscillatory or slowly varying processes, and defined for such processes the evolutionary spectral density function which has physical properties similar to those of a stationary second order spectrum. Priestley and Gabr [15] then extended this concept to the case of evolutionary bispectra, and considered the distribution and sampling properties of evolutionary bispectral estimates.

In this paper we use the concept of evolutionary bispectrum in order to construct statistical tests for Gaussianity and linearity of nonstationary slowly varying processes. In order to do this, we begin by giving a brief outline in section 2 of the existing theory on the evolutionary second order

spectrum and bispectrum in which oscillatory and linear oscillatory processes are defined. In section 3 we review some of the existing theory on estimating the evolutionary spectrum and bispectrum. In section 4 we develop tests for Gaussianity and linearity for oscillatory processes based on the theory given in sections 2 and 3. These provide a generalization to earlier tests on Gaussianity and linearity (see [21]) for stationary processes. In section 5 we consider a time series of network measurements arising from internet traffic.

Several models have been proposed in the past in describing internet traffic. However, such models exhibit stationarity which in many cases was found to be unrealistic for such processes (see e.g. [4], [28]). It has also been suggested that internet traffic data may be described by a model whose parameters vary with time (see [10]). It would therefore be reasonable to study internet traffic data under the assumption that the nonstationarity is one exhibited by a slowly varying process. Furthermore, Igloi and Terdik [7], Terdik and Molnar [24] and Molnar and Terdik [9] proposed new nonfractal models which are not self similar and the scaling factor can be estimated using higher order spectral estimates. In addition, Basu, Mukherjee and Klivansky [1] proposed non-Gaussian nonstationary models for internet traffic which after differencing become stationary. It would thus be interesting to investigate whether a Gaussian/linear class of nonstationary models which do not admit a simple differencing strategy and are slowly varying would be

appropriate in describing internet traffic data.

2 Evolutionary second order spectrum and bispectrum

In this section we briefly review some of the results given in [12] and [15].

Let $\{X_t\}$ be a zero mean discrete parameter time series admitting a representation

$$X_t = \int_{-\pi}^{\pi} e^{it\omega} A_t(\omega) dZ(\omega)$$

where $Z(\omega)$ is an orthogonal random process, with

$$E[dZ(\omega)] = 0,$$

$$E[|dZ(\omega)|^2] = d\mu(\omega)$$

and for each fixed ω, $A_t(\omega)$ has a Fourier transform whose absolute maximum occurs at the origin. Priestley [12] defines such a process $\{X_t\}$ an 'oscillatory process'.

The evolutionary spectral density function $h_t(\omega)$, and the evolutionary bispectral density function $h_t(\omega_1, \omega_2)$, are then defined by

$$\begin{aligned} h_t(\omega)d(\omega) &= |A_t(\omega)|^2 d\mu(\omega), \\ h_t(\omega_1, \omega_2)d\omega_1 d\omega_2 &= A_t(\omega_1)A_t(\omega_2)A_t(-\omega_1 - \omega_2)d\mu(\omega_1, \omega_2), \end{aligned}$$

where

$$d\mu(\omega) = E[|dZ(\omega)|^2],$$

$$d\mu(\omega_1, \omega_2) = E[dZ(\omega_1)dZ(\omega_2)dZ(-\omega_1 - \omega_2)].$$

It is shown in [15] that if a nonstationary time series $\{X_t\}$ is Gaussian then $h_t(\omega_1, \omega_2) = 0$ for all t, ω_1 and ω_2.

Let us now describe some properties of the evolutionary spectrum and bispectrum (for details see [17], [23], [19] and [20]). Let $\{Y_t\}$ and $\{X_t\}$ be two zero mean oscillatory processes and let

$$X_t = \sum_{u=0}^{\infty} g_{t,u} Y_{t-u}$$

where the filter $\{g_{t,u}\}$ is a deterministic function of u and t. Let also $\Gamma_t(\omega) = \sum_{u=0}^{\infty} g_{t,u} e^{-i\omega u}$. Under suitable conditions, Priestley and Gabr [15] showed that

$$\begin{aligned} h_{t,X}(\omega) &\sim |\Gamma_t(\omega)|^2 h_{t,Y}(\omega), \\ h_{t,X}(\omega_1, \omega_2) &\sim \Gamma_t(\omega_1)\Gamma_t(\omega_2)\Gamma_t(-\omega_1 - \omega_2) h_{t,Y}(\omega_1, \omega_2) \end{aligned} \tag{1}$$

where $|\omega| \leq \pi$, $\omega_1 + \omega_2 + \omega_3 = 0 \ (mod\ 2\pi)$, $h_{t,X}(\omega)$ and $h_{t,X}(\omega_1, \omega_2)$ are respectively the evolutionary spectrum and the evolutionary bispectrum of $\{X_t\}$. It is important to note the remarkable similarity between the relations (1) and those that exist for stationary processes.

2.1 Linear Processes

An oscillatory process $\{X_t\}$ is said to be a linear process if it admits the representation

$$X_t = \sum_{u=0}^{\infty} g_{t,u} e_{t-u} \tag{2}$$

where $\{e_t\}$ are independent, identically distributed random variables with $E[e_t] = 0$, $E[e_t^2] = \sigma_e^2$ and $E[e_t^3] = \mu_3$. If $\{X_t\}$ admits the representation (2), then we have

$$h_{t,X}(\omega) = \frac{\sigma_e^2}{2\pi} |\Gamma_t(\omega)|^2, \tag{3}$$

$$h_{t,X}(\omega_1, \omega_2) = \frac{\mu_3}{(2\pi)^2} \Gamma_t(\omega_1)\Gamma_t(\omega_2)\Gamma_t(-\omega_1 - \omega_2). \tag{4}$$

If $\{e_t\}$ is Gaussian then $\mu_3 = 0$, which implies that $h_{t,X}(\omega_1, \omega_2) = 0$ for all ω_1, ω_2 and t. More generally, we obtain from (1) and (4),

$$|h_{t,X}(\omega_1, \omega_2)|^2 = \frac{\mu_3^2}{2\pi\sigma_e^6} h_{t,X}(\omega_1) h_{t,X}(\omega_2) h_{t,X}(-\omega_1 - \omega_2)$$

which implies that

$$\frac{|h_{t,X}(\omega_1, \omega_2)|^2}{h_{t,X}(\omega_1) h_{t,X}(\omega_2) h_{t,X}(-\omega_1 - \omega_2)} = \frac{\mu_3^2}{2\pi\sigma_e^6} \tag{5}$$

and the ratio on the left is thus independent of ω_1, ω_2 and t. In other words, in testing the Gaussianity of a nonstationary series we can test the null hypothesis $h_t(\omega_1, \omega_2) = 0$ for all ω_1 and ω_2. Also in testing for linearity we can test for the constancy of the ratio (5) using a procedure similar to that

described in Subba Rao and Gabr [22]. It may be pointed out that in testing for second order stationarity, i.e. testing $h_t(\omega) = h(\omega)$ for all t, we use the statistical tests proposed in [16]. A similar test in testing for stationarity of higher order moments can also be constructed. In the following, we assume that the series is nonstationary and we construct tests for Gaussianity and linearity.

3 Estimation of the evolutionary spectrum and bispectrum

Let us now consider briefly the estimation of spectra and bispectra (for details see [12] and [15]). Let $(X_1, X_2, \ldots, X_N)$ be a sample from a zero mean discrete parameter nonstationary time series $\{X_t\}$. Let $\{g_u\}$ be a filter and $\Gamma(\omega)$ the corresponding frequency response function, i.e.

$$\Gamma(\omega) = \sum_u g_u e^{-i\omega u}.$$

The bandwidth of $\{g_u\}$ is defined as $B_g = \sum_u |u||g_u|$ and we assume that $\{g_u\}$ is normalised so that

$$2\pi \sum_{u=-\infty}^{\infty} |g_u|^2 = \int_{-\pi}^{\pi} |\Gamma(\omega)|^2 d\omega = 1,$$

and

$$\sum_{u=-\infty}^{\infty} |u||g_u| = B_g.$$

Now write

$$U_t(\omega_0) = \sum_{u=-\infty}^{\infty} g_u X_{t-u} e^{-i\omega_0(t-u)}.$$

We may then construct evolutionary spectral estimates using

$$\hat{h}_t(\omega) = \sum_{v=-\infty}^{\infty} w_{T',v} |U_{t-v}(\omega)|^2$$

where the weight function $w_{T',t}$ satisfies the following conditions:

- $w_{T',t} \geq 0$ for all t, T'
- $w_{T',t}$ decays to zero as $|t| \to \infty$, for all T'
- $\sum_{t=-\infty}^{\infty} w_{T',t} = 1$, for all T'
- $\sum_{t=-\infty}^{\infty} \{w_{T',t}\}^2 < \infty$, for all T'.

If we assume that $\{g_u\}$ is normalized so that

$$\int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \Gamma(\omega_1)\Gamma(\omega_2)\Gamma(-\omega_1 - \omega_2) d\omega_1 d\omega_2 = 1$$

we may also construct evolutionary bispectral estimates using

$$\hat{h}_t(\omega_1, \omega_2) = \sum_{v=-\infty}^{\infty} w_{T',v} U_{t-v}(\omega_1) U_{t-v}(\omega_2) U_{t-v}(-\omega_1 - \omega_2)$$

where the weight function $w_{T',t}$ satisfies the same conditions as above.

Let $W_{T'}(\lambda) = \sum_{t=-\infty}^{\infty} e^{-i\lambda t} w_{T',t}$ and assume there is a constant C such that

$$\lim_{T' \to \infty} \left\{ T' \int_{-\pi}^{\pi} |W_{T'}(\theta)|^2 d\theta \right\} = C.$$

Under the above conditions on the filter $\{g_u\}$ and the weight function $w_{T',v}$ Priestley [14] has shown that

$$E\{\hat{h}_t(\omega)\} \sim \int_{-\pi}^{\pi} \bar{h}_t(\omega + \omega_0) |\Gamma(\omega)|^2 d\omega$$

and

$$T' Var\{\hat{h}_t(\omega_0)\} \sim (1 + \delta_{0,\omega_0}) C \tilde{h}_t^2(\omega_0) \left\{ \int_{-\pi}^{\pi} |\Gamma(\theta)|^4 d\theta \right\}$$

where

$$\bar{h}_t(\omega) = \sum_{v=-\infty}^{\infty} w_{T',v} h_{t-v}(\omega),$$

$$\tilde{h}_t^2(\omega_0) = \frac{\sum_{v=-\infty}^{\infty} h_{t-v}^2(\omega_0) \{w_{T',v}\}^2}{\sum_{v=-\infty}^{\infty} \{w_{T',v}\}^2}$$

and $\delta_{.,.}$ is the Kronecker delta function. Note that the sampling properties of $\hat{h}_t(\omega)$ have been stated in [14]. Although no rigorous proof exists on the asymptotic results given, a heuristic argument in the same paper shows that these properties should reasonably well be expected to hold. Furthermore, it has been demonstrated by Tsolaki in [25] using Monte Carlo methods that $\hat{h}_t(\omega)$ has approximately a normal distribution and that a logarithmic transformation of $\hat{h}_t(\omega)$ makes the estimate converge to a normal random variable faster, as conjectured by Priestley and Subba Rao in [16].

In estimating $h_t(\omega)$ a choice of windows $\{g_u\}$ and $w_{T',v}$ is required. For a discussion on different choices of windows we refer to Priestley [13]. We note that the computations in section 5 were carried out for various choices of $\{g_u\}$ and $w_{T',v}$ all of them supporting our conclusions. In this paper we present the results obtained for

$$g_u = \begin{cases} \frac{1}{\sqrt{2\pi(2h+1)}}, & |u| \leq h, \\ 0, & otherwise \end{cases} \tag{6}$$

and

$$w_{T',v} = \begin{cases} \frac{1}{T'+1}, & -\frac{T'}{2} \leq v \leq \frac{T'}{2}, \\ 0, & otherwise. \end{cases} \tag{7}$$

We now consider the sampling properties of $\hat{h}_t(\omega_1, \omega_2)$. Priestley and Gabr [15] have shown that

$$E[\hat{h}_t(\omega_1, \omega_2)] \sim$$
$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \overline{h}_t(\omega_1 + v_1, \omega_2 + v_2)\Gamma(v_1)\Gamma(v_2)\Gamma(-v_1 - v_2)dv_1 dv_2$$

where

$$\overline{h}_t(\omega_1, \omega_2) = \sum_{u=-\infty}^{\infty} w_{T',v} h_{t-v}(\omega_1, \omega_2)$$

and

$$Var[\hat{h}_t(\omega_1, \omega_2)] \sim 2\pi[\delta(\omega_2)\{1 + 8\delta(\omega_1)\}G_1$$

$$+\{1+\delta(\omega_1-\omega_2)+4\delta(\omega_1)\delta(\omega_2)\}G_2]$$
$$\times\left\{\sum_{v=-\infty}^{\infty} w_{T',v}^2\right\}\tilde{h}_t(\omega_1,\omega_2)$$

where $\delta_{.,.}$ is the Kronecker delta and

$$\tilde{h}_t(\omega_1,\omega_2)=\frac{\sum_{u=-\infty}^{\infty}\{h_{t-u}(\omega_1)h_{t-u}(\omega_2)h_{t-u}(\omega_1+\omega_2)\}w_{T',u}^2}{\sum_{v=-\infty}^{\infty} w_{T',v}^2},$$

$$G_1=|\Gamma(0)|^2\left\{\int_{-\pi}^{\pi}|\Gamma(\omega)|^2d\omega\right\}^2,$$

$$G_2=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\{\Gamma(\omega_1)\Gamma(\omega_2)\Gamma(-\omega_1-\omega_2)\}^2d\omega_1 d\omega_2.$$

Assuming that $h_t(\omega_1,\omega_2)$ is 'smooth' with respect to $\Gamma(\omega_1)\Gamma(\omega_2)\Gamma(-\omega_1-\omega_2)$ and that the bandwidth of $w_{T',v}$ is 'small' compared with the 'time domain bandwidth' of $h_t(\omega_1,\omega_2)$ we have that

$$E[\hat{h}_t(\omega_1,\omega_2)]\sim h_t(\omega_1,\omega_2)$$

and

$$Var[\hat{h}_t(\omega_1,\omega_2)]\sim 2\pi[\delta(\omega_2)\{1+8\delta(\omega_1)\}G_1$$
$$+\{1+\delta(\omega_1-\omega_2)+4\delta(\omega_1)\delta(\omega_2)\}G_2]$$
$$\times\left\{\sum_{v=-\infty}^{\infty} w_{T',v}^2\right\}h_t(\omega_1)h_t(\omega_2)h_t(\omega_1+\omega_2)$$
$$=B(\omega_1,\omega_2)h_t(\omega_1)h_t(\omega_2)h_t(\omega_1+\omega_2)$$

where

$$B(\omega_1,\omega_2)=2\pi[\delta(\omega_2)\{1+8\delta(\omega_1)\}G_1+$$
$$+\{1+\delta(\omega_1-\omega_2)+4\delta(\omega_1)\delta(\omega_2)\}G_2]\times\left\{\sum_{v=-\infty}^{\infty} w_{T',v}^2\right\}. \qquad (8)$$

We also note that the bispectrum $h_t(\omega_1, \omega_2)$, which is complex valued, satisfies the usual symmetry relations

$$h_t(\omega_1, \omega_2) = h_t(\omega_2, \omega_1) = h_t(\omega_1, -\omega_1 - \omega_2) =$$

$$= h_t(-\omega_1 - \omega_2, \omega_2) = h_t^*(-\omega_1, -\omega_2)$$

where $-\pi \leq \omega_1,\ \omega_2 \leq \pi$. As a result it suffices to consider only frequencies in the domain

$$\Omega = \{(\omega_1, \omega_2);\ 0 \leq \omega_2,\ \omega_2 \leq \omega_1,\ \omega_1 + \omega_2 \leq \pi\}. \tag{9}$$

One can see from the definition of the evolutionary bispectrum that $h_t(\omega_1, \omega_2)$ is real valued at $\omega_1 = 0$, $\omega_2 = 0$ as well as along the boundaries.

In estimating $h_t(\omega_1, \omega_2)$ we use

$$g_u = \begin{cases} \frac{1}{\sqrt[3]{4\pi^2(2h+1)}}, & |u| \leq h, \\ 0, & otherwise. \end{cases} \tag{10}$$

The weight function $w_{T',v}$ is chosen to be of the form (7). We note that $\{g_u\}$ as given in (10) is different to the one used for the estimation of the spectral density function as it is chosen to satisfy a different normalizing condition.

Using Monte Carlo methods, Tsolaki [25] has shown further that there is strong evidence to support that the distribution of the evolutionary bispectral estimate is approximately complex Gaussian.

4 Tests for Gaussianity and linearity

As already mentioned in section 2, if a series is Gaussian, its evolutionary bispectral density function $h_t(\omega_1, \omega_2)$ is zero for all values of ω_1, ω_2 and for all values of t. The statistic we propose should depend on all frequencies and time points. However, in view of the symmetry relations we can restrict our frequency domain to the principal domain given in (9).

We take the bispectral estimate $\hat{h}_t(\omega_1, \omega_2)$ to be approximately distributed as complex normal with mean $h_t(\omega_1, \omega_2)$ and variance $B(\omega_1, \omega_2)h_t(\omega_1)h_t(\omega_2)h_t(\omega_1 + \omega_2)$. It follows that

$$X_t(\omega_1, \omega_2) = \frac{2[(\hat{h}_t(\omega_1, \omega_2))^2]}{B(\omega_1, \omega_2)h_t(\omega_1)h_t(\omega_2)h_t(\omega_1 + \omega_2)}$$

is distributed approximately as a noncentral χ^2 with 2 degrees of freedom with noncentrality parameter

$$\lambda_t(\omega_1, \omega_2) = \frac{2\,|h_t(\omega_1, \omega_2)|^2}{B(\omega_1, \omega_2)h_t(\omega_1)h_t(\omega_2)h_t(\omega_1 + \omega_2)}.$$

4.1 Test for Gaussianity

If $h_t(\omega_1, \omega_2) = 0$ then $X_t(\omega_1, \omega_2)$ is distributed as a central χ^2 with 2 degrees of freedom. As an overall measure for departure of Gaussianity we consider the statistic

$$Y_1 = \sum_{i=1}^{I} \sum_{\omega_1,\, \omega_2 \in G} X_{t_i}(\omega_1, \omega_2)$$

where the time points $t_1, t_2, \ldots, t_I$ and the frequency points G are chosen such that the statistics $X_{t_i}(\omega_1, \omega_2)$ are approximately independent (see [15]). If the set G contains J points, then under the null hypothesis Y_1 will be distributed as a central χ^2 with $2IJ$ degrees of freedom.

In the computation of the statistic $X_t(\omega_1, \omega_2)$ the assumption that the evolutionary spectral density function $h_t(\omega)$ is known is unrealistic. In practice we assume that the spectral density function $h_t(\omega)$ is estimated following the procedure suggested by Priestley [12]. The resulting test statistic will still be approximately distributed as a central χ^2 under the null hypothesis. The statistic Y_1 does not contain $X_t(\omega_1, \omega_2)$ estimated along the boundary $\omega_2 = 0$. We note that on the boundary the estimated evolutionary bispectrum $\hat{h}_t(\omega_1, \omega_2)$ is normal. At these frequencies the test statistic $X_t(\omega_1, \omega_2)$ under the null hypothesis is a central χ^2 with 1 degree of freedom. Taking into account the values on the boundary we consider the statistic

$$Y_2 = \sum_{i=1}^{I} \sum_{\omega_2=0,\ all\ \omega_1} \frac{|\hat{h}_{t_i}(\omega_1, \omega_2)|^2}{B(\omega_1, \omega_2)\hat{h}_t(\omega_1)\hat{h}_t(\omega_2)\hat{h}_t(\omega_1 + \omega_2)}$$

which under the null hypothesis is distributed as a central χ^2 distribution with $J'I$ degrees of freedom where J' is the number of frequencies considered on the line $\omega_2 = 0$. As an overall measure of departure from Gaussianity we consider the test statistic $\Delta_1 = Y_1 + Y_2$ where Y_1 and Y_2 are independent. Under the null hypothesis, Δ_1 is distributed as a central χ^2 with $I(2J + J')$

degrees of freedom. If $\Delta_1 > \chi^2_{\alpha,I(2J+J')}$ we conclude that there is evidence to suggest that the nonstationary signal is non Gaussian.

4.2 Test for Linearity

As pointed out earlier, nonstationary series can be non Gaussian but linear in the sense that $\{X_t\}$ admits the linear representation (2). Under the hypothesis that the ratio (5) is independent of ω_1, ω_2 and t, $\hat{h}_t(\omega_1, \omega_2)$ is approximately normally distributed with mean $h_t(\omega_1, \omega_2)$ and variance $B(\omega_1, \omega_2)h_t(\omega_1)h_t(\omega_2)h_t(\omega_1 + \omega_2)$. Hence under the null hypothesis that the time series is linear, the statistic

$$Z_t(\omega_1, \omega_2) = \frac{\sqrt{2}|\hat{h}_t(\omega_1, \omega_2)|}{[B(\omega_1, \omega_2)\hat{h}_t(\omega_1)\hat{h}_t(\omega_2)\hat{h}_t(\omega_1 + \omega_2)]^{\frac{1}{2}}}$$

is approximately normally distributed with mean $c\sqrt{\frac{2}{B(\omega_1,\omega_2)}}$ and variance 1 where $c = \frac{|\mu_3|}{\sqrt{2\pi}\sigma_e^3}$. As c is unknown, the mean of $Z_t(\omega_1, \omega_2)$ is unknown but an estimate of the mean can be found using

$$\left|\overline{\hat{h}}(\omega_1, \omega_2)\right| = \frac{1}{NM\left\{\frac{B(\omega_1,\omega_2)}{2}\right\}^{\frac{1}{2}}} \sum_{all\ t} \sum_{\omega_1,\omega_2} \frac{\left|\hat{h}_t(\omega_1, \omega_2)\right|}{\left\{\hat{h}_t(\omega_1)\hat{h}_t(\omega_2)\hat{h}_t(\omega_1 + \omega_2)\right\}^{\frac{1}{2}}}$$

where $B(\omega_1, \omega_2)$ is a deterministic function and can be calculated using (8). Hence, asymptotically

$$\left(Z_t(\omega_1, \omega_2) - \left|\overline{\hat{h}}(\omega_1, \omega_2)\right|\right) \sim N(0, 1)$$

and therefore, the statistic

$$Y_3 = \sum_{all\ t} \sum_{\omega_1,\omega_2} \left(Z_t(\omega_1,\omega_2) - \left|\overline{\hat{h}}(\omega_1,\omega_2)\right| \right)^2 \sim \chi^2_{IJ}(0) \qquad (11)$$

is distributed (under the null hypothesis of linearity) approximately as a central χ^2 with IJ degrees of freedom.

In the evaluation of the summation (11) we have not considered the boundary $\omega_2 = 0$ where at these frequencies (i.e. $\omega_2 = 0$ for all ω_1) the statistic

$$Z_t(\omega_1,\omega_2) = \frac{\hat{h}_t(\omega_1,\omega_2)}{(B(\omega_1,\omega_2)h_t(\omega_1)h_t(\omega_2)h_t(\omega_1+\omega_2))^{1/2}}$$

is approximately normal with mean $\frac{c}{(B(\omega_1,\omega_2))^{\frac{1}{2}}}$ and variance 1. Hence we consider the statistic

$$Y_4 = \sum_t \sum_{\omega_1} (Z_t(\omega_1,0) - \bar{Z}(\omega_1,0))^2$$

where

$$\bar{Z}(\omega_1,0) = \frac{1}{IJ'\{B(\omega_1,0)\}^{\frac{1}{2}}} \sum_t \sum_{\omega_1} \frac{\hat{h}_t(\omega_1,0)}{(\hat{h}_t(\omega_1)\hat{h}_t(0)\hat{h}_t(\omega_1))^{1/2}}.$$

Under the null hypothesis, Y_4 is distributed as a central χ^2 with IJ' degrees of freedom. As an overall measure of departure from linearity we consider the statistic $\Delta_2 = Y_3 + Y_4$ where Y_3 and Y_4 are independent and is distributed as a central χ^2 with $I(J+J')$ degrees of freedom. We reject the null hypothesis if $\Delta_2 > \chi^2_{\alpha,I(J+J')}$.

5 An application to internet traffic data

Management of internet traffic has become one of the most important tasks in present day communication networks. An understanding of internet traffic can contribute to fast and efficient communications and thus suitable statistical models are needed to describe its complex structure. These models can be used to simulate internet traffic. In a recent excellent review, Cleveland and Sun [5] provide some basic ideas on internet traffic modelling as well as a statistical analysis of the traffic. In internet communications, information is transferred from one computer to another using the *IP* (Internet Protocol) which implements two basic functions: addressing and fragmentation, examples of which are webpages and e-mails. The transfer is carried out by different protocols based on the application. For example, the *HTTP* (Hypertext Transfer Protocol) transfers a worldwide webpage from a server computer to a client computer, *SMTP* (Simple Mail Transfer Protocol) sends e-mails and *FTP* (File Transfer Protocol) transfers files between local and remote network computers.

In an *ATM* (Asynchronous Transfer Mode) network when files are transferred they are divided into packets which are reassembled at the receiver end of the computer. Each packet consists of static capacity which in our implementation was 1460 bytes of information. Each computer has a unique

Internet Protocol (*IP*) number which is carried by each packet when it is sent. A computer connects with another using the Transfer Control Protocol (*TCP*).

Ethernet works at layer 1-the Physical Layer. This means that any protocol can be used, TCP/IP, NetBeui, IPX. However, if two sources are trying to send simultaneously packets to destinations, there is a mechanism called Carrier Sense Multiple Access with Collision Detection (CSMA/CD) which is a non-deterministic Media Access Control (MAC) protocol in which a node verifies the absence of other traffic on a shared physical medium before transmitting. This mechanism is soundly responsible to avoid collisions by transmitting a short signal to indicate a node's intention to transmit. When other nodes see this signal, they wait for some time before attempting to send any frames. In this way collision can be avoided on any Ethernet-based access media.

Internet traffic can be viewed as a point process (arrival times of packets). If the size of packets (in bytes) is sent together with arrival times then the series is a marked point process. When byte counts are summed over equally spaced time intervals, a time series is obtained. In this paper we consider such a time series for our analysis. Cleveland and Sun [5] have pointed out that *HTTP* start times are nonstationary and nonstationarity is as pervasive as internet traffic data is long range persistent. This view is confirmed in the

following analysis using evolutionary spectral methods.

The data we consider is 425 million arrivals (at workstations) which is the number of nonempty packets arriving at a workstation, collected at the Swedish University Network. In experiments, every cluster of workstations has its particular specified mean response time depending on CPU processing/response time and network metrics as well as on network dimensions particularly when an wide area network is considered. However, in discrete time the sampling interval must be relatively small in order to avoid the self-similarity nature of traffic data. In our case, the number of nonempty packets arriving at a workstation were summed over 20 ms time intervals. The data were divided into 8000 slots where each block consists of 40000 to 60000 measurements. In this paper we consider one such block consisting of 53136 measurements. For the estimation of the evolutionary spectrum and bispectrum we have used several choices of h and T'. Depending on the choices of h and T' in estimating the spectrum and bispectrum we need to choose the time and frequency points to be sufficiently apart so that the estimates are approximately uncorrelated (see [12]). Here we present the analysis made for a choice of $h = 7$ and $T' = 300$ in estimating the spectral density function and $h = 4$, $T' = 120$ in estimating the bispectrum.

Plots of parts of the data are shown in figure (1) corresponding to four segments of the data those at time points $1 - 1000$, $15001 - 16000$, $30001 -$

31000 and 45001 − 46000 respectively.

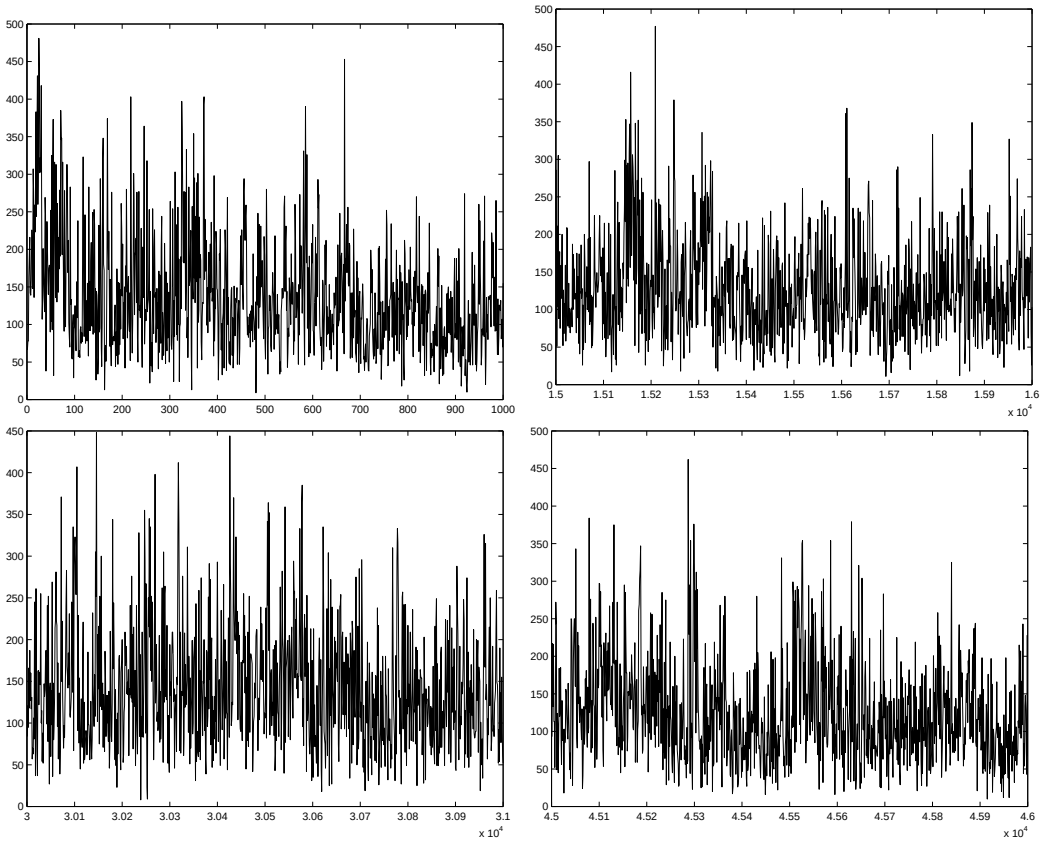

Figure 1: Plot of segments of the data corresponding to points 1-1000, 15001-16000, 30001-31000 and 45001-46000 respectively.

As we cannot include a plot of the entire series we have chosen four segments of the data which include some of the time points at which we estimate the spectral density function and the bispectrum. The test of Gaussianity and linearity is applied to the data after we remove the trend; we first remove a constant mean (denoted by series c), then a linear trend (denoted by series l) and finally a quadratic trend (denoted by series q). As the results are similar for the three data sets the figures presented in the subsequent of the paper are those for the data the we have removed the constant mean.

Furthermore, it is not clear whether we can assume that the variance is constant. As a change in the covariance structure of a series reflects to a change in the spectrum, we test the data for stationarity by applying the test by Priestley and Subba Rao [16]. We have applied the tests to all three series to see if a possible nonstationarity in the mean would affect the conclusions of the test. A plot of the evolutionary spectrum at time points 500, 10500, 20500, 30500, 40500 and 50500 for data series (c) are shown in figure (2). The test of nonstationarity shows that there is evidence to reject the null hypothesis that the series is stationary for all three data sets. Thus, we procceed in applying the test of Gaussianity and linearity to the three data sets.

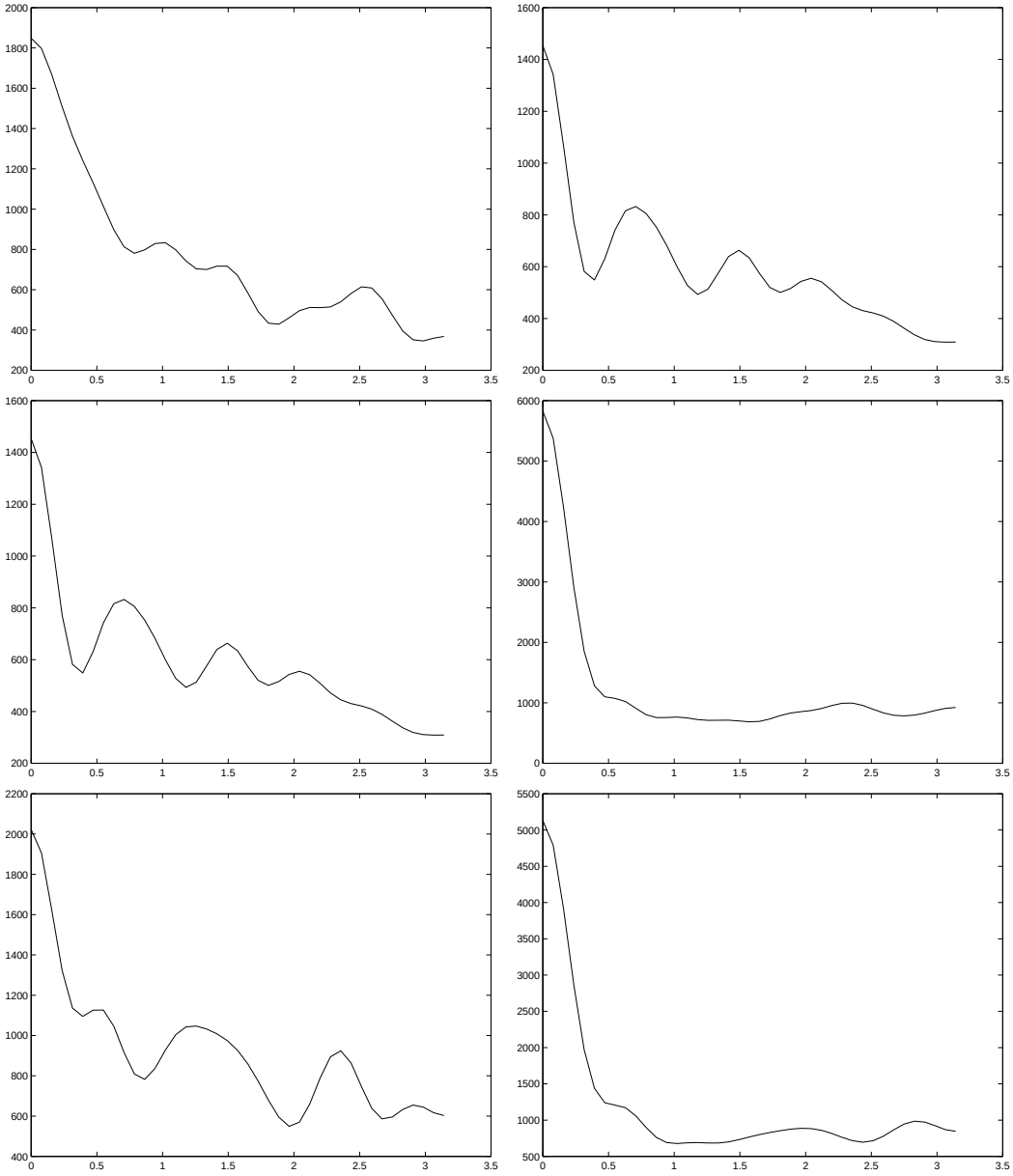

Figure 2: Plot of the estimated spectral density function of the internet data at time points 500, 10500, 20500, 30500, 40500 and 50500.

In order to test the series for Gaussianity and linearity the evolutionary bispectrum and normalized bispectrum are estimated at time points (500(5000)53136) and frequency points $(\frac{\pi}{6}(\frac{\pi}{6})\frac{5\pi}{6})$. Plots of the bispectrum and normalized bispectrum of series (c) for time points 500, 10500, 20500, 30500, 40500 and 50500 are shown in figures (3) and (4) respectively. The estimates show differences in average magnitudes suggesting possibly non Gaussianity and nonlinearity in the data. To test such a hypothesis we apply the proposed tests.

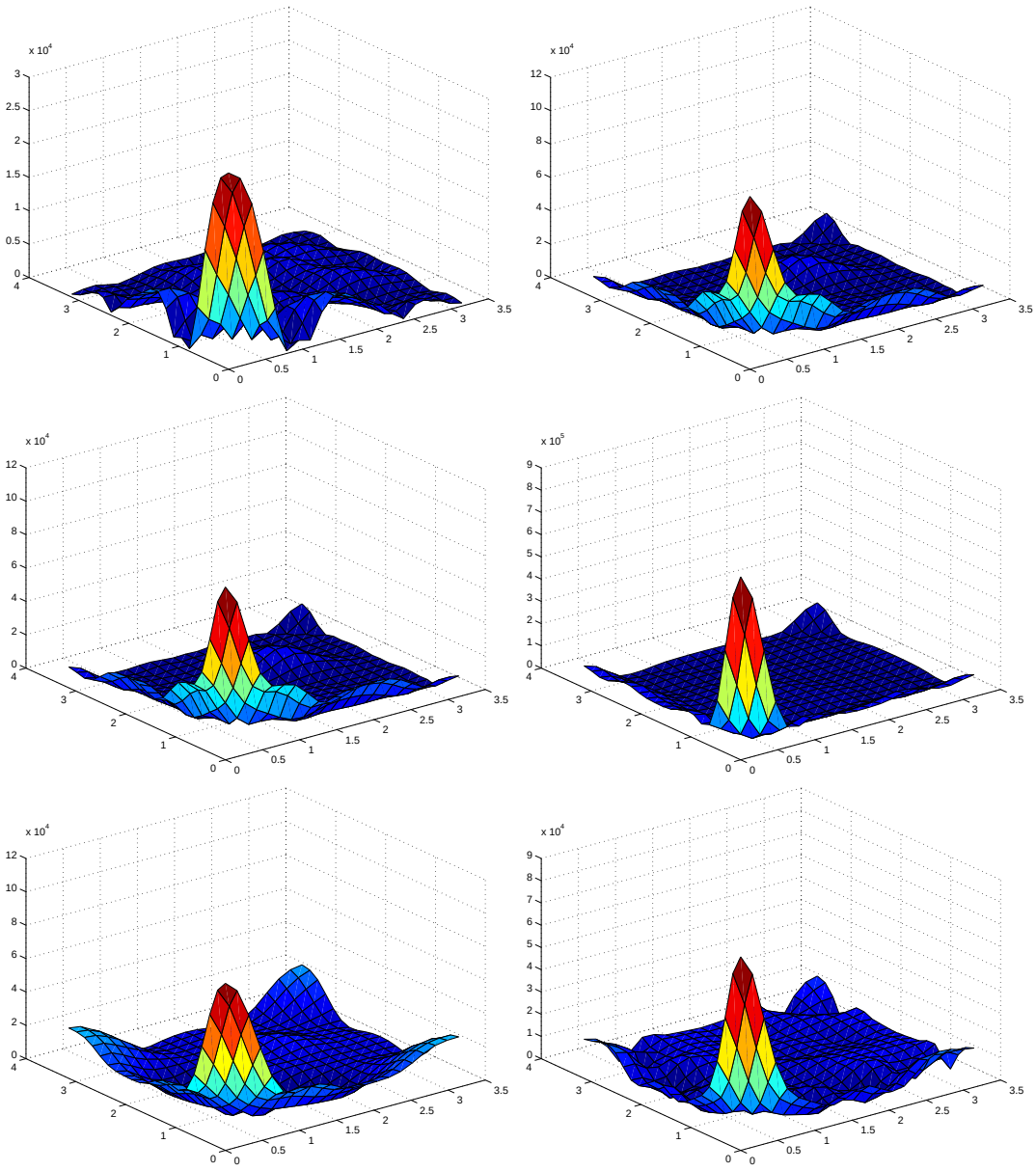

Figure 3: Plot of the estimated bispectrum of the internet data at time points 500, 10500, 20500, 30500, 40500 and 50500.

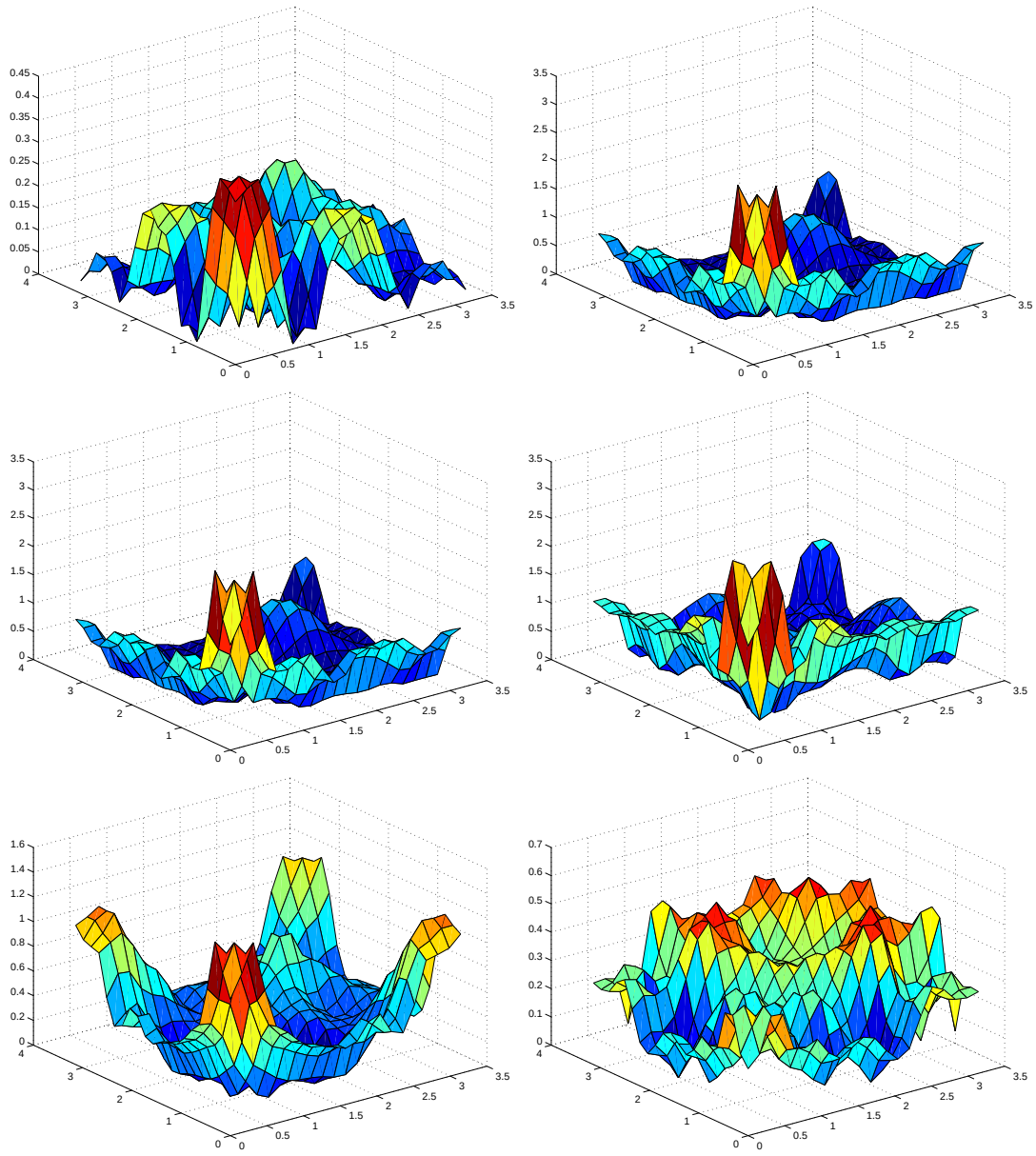

Figure 4: Plot of the estimated normalized bispectrum of the internet data at time points 500, 10500, 20500, 30500, 40500 and 50500.

We calculate the test statistics Δ_1 and Δ_2 following the procedure described in sections (4.1) and (4.2). The value of the test statistic Δ_1 is 774.4 for series (c), 765.7 for series (l) and 776.8 for series q which is compared with the $\chi^2_{0.05}$ with 341 degrees of freedom which is 385.1. The degrees of freedom are so large, one could use critical points from a Normal distribution. The test statistic for all three series is much higher than the critical value and thus we can conclude that there is evidence to suggest that the data is non-Gaussian. Also we compare the values of the test statistic Δ_2 which is 290.1 for series (c), 278.5 for series (l) and 292.6 for series (q) with the $\chi^2_{0.05}$ with 209 degrees of freedom. The value of the test statistic is higher than the critical value (243.8) and thus the hypothesis of linearity is rejected at 5% significance level for the three series. Thus the Internet Traffic data are nonstationary non-Gaussian and nonlinear.

Our conclusions of the data being nonstationary drawn using evolutionary spectral methods, confirm the hypothesis of Cao, Cleveland, Lin and Sun [4]. The reason of nonstationarity in the internet traffic is believed to have been caused by the superposition of internet traffic from various sources such as *HTTP*, *FTP* and telnet (Cleveland and Sun [5]). Recent studies by Terdik and Molnar [24], Igloi and Terdik [7] and Basu, Mukherjee and Klivansky [1] indicated that internet traffic data is non Gaussian which is also confirmed by this analysis. In addition it is shown here that the data is not only

nonstationary and non Gaussian but also nonlinear. Therefore, in order to model the data, a suitable nonstationary and nonlinear time series model must be found. If a specific form of a nonlinear process is assumed then one could estimate a linear representation (see e.g. [6]) for such a process.

Acknowledgment

The authors would like to thank Professors S. Molnar and Gy. Terdik for allowing us to use their internet traffic data.

References

[1] Sabyasachi Basu, Amarnath Mukherjee, Steve Klivansky, *Time Series Models for Internet Traffic*, Technical Report GIT-CC-95-27, Georgia Institure of Technology, 1996.

[2] David R. Brillinger, An Introduction to Polyspectra, *Annals of Mathematical Statistics*, 36, 1351-1374, 1965.

[3] David R. Brillinger, Murray Rosenblatt, Asymptotic theory of estimates of k-th order spectra, *Spectral Analysis Time Series* (Proc. Advanced Sem., Madison, Wis., 1966), 153-188, John Wiley, New York, 1967.

[4] Jin Cao, William S. Cleveland, Dong Lin and Don X. Sun, *On the Nonstationarity of Internet Traffic*, ACM SIGMETRICS, 102-112, 2001

[5] William S. Cleveland and Don X. Sun, *Internet Traffic Data*, Journal of the American Statistical Association, 95, 979-985, 2000. Reprinted in Statistics in the 21st Century, 214-228, edited by A. E.

[6] Christian Francq, Jean-Michel Zakoïan, *Estimating linear representations of nonlinear processes*, Journal of statistical planning and inference, 68, 145-165, 1998.

[7] Endre Igloi, Gyorgy Terdik, *Superposition of diffusions with linear generator and its multifractal limit process*, ESAIM Prob. Stat. 7, 23-88, 2003.

[8] K. S. Lii, M. Rosenblatt, Deconvolution and Estimation of Transfer Function Phase and Coefficients for Non-Gaussian Linear Processes, *Ann. Statist.*, 20, 1195-1208, 1982.

[9] S. Molnar, Gy. Terdik, *A Monofractal Model for Network Traffic*, Technical Report, COST 279TD(02)04, Leidschendam, The Netherlands, February 7-8, 2002.

[10] A. Mukherjee, *On the dynamics and significance of low frequency components of internet load*, Technical Report MS-CIS-92-83/DSL-12, Computer and Information Science Department, University of Pennsylvania.

[11] C. L. Nikias, A. P. Petropulu, *Higher-Order Spectral Analysis: A Nonlinear Signal Processing Framework*, New Jersey: Prentice-Hall, 1993.

[12] M. B. Priestley, Evolutionary spectra and non-stationary processes, *J. R. Statistical Society, Series B*, 27, 204-237, 1965.

[13] M. B. Priestley, *Spectral Analysis and Time Series*, 2 vols. Academic Press, London and New York, 1981.

[14] M. B. Priestley, Design relations for non-stationary processes, *J. Roy. Statist. Soc. Ser. B*, 28, 228-240, 1966.

[15] M. B. Priestley, M. M. Gabr, Bispectral analysis of non-stationary processes. In *Multivariate Analysis: Future Directions.* edited by C. R. Rao, Elsevier Science Publishers, Chapter 16 , 295-317, 1993.

[16] M. B. Priestley, T. Subba Rao, A test for stationarity of time series. *J. Roy. Statist. Soc. Ser. B*, 31, 140-149, 1969.

[17] M. B. Priestley, H. Tong, On the analysis of bivariate non-stationary processes. With comments by M. S. Bartlett, A. M. Walker, T. Subba Rao, M. D. Godfrey, W. D. Ray, H. E. Daniels, J. Durbin, and a reply by M. B. Priestly and H. Tong and by J. K. Hammond. *J. Roy. Statist. Soc. Ser. B*, 35, 153-166, 179-188, 1973.

[18] M. Rosenblatt, J. W. Van Ness, Estimation of the bispectrum, *Ann. Math. Statist.*, 36 1965 1120–1136.

[19] T. Subba Rao, Discussion of the paper by Professor Priestley and Dr Tong and of the paper by Dr Hammond. *J. R. Statist. Soc. B*, 31, 140, 1973.

[20] T. Subba Rao, Statistical analysis of nonlinear and nonGaussian time series. In *Stochastic differential and difference equations*, edited by Csiszar Gy Michaletzky, Birkhauser, Boston, 285-298, 1997.

[21] T. Subba Rao, M. M. Gabr, A test for linearity of stationary time series. *J. Time Series Anal.*, 1, 145-158, 1980.

[22] T. Subba Rao, M. M. Gabr, *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Springer-Verlag, Berlin, 1984.

[23] T. Subba Rao, H. Tong, A Test for Time-Dependence of Linear Open-loop Systems. *The Journal of the Royal Statistical Society, Series B (Methodological)*, 34, No. 2, 235-250, 1972.

[24] Gy. Terdik, S. Molnar, *A General Fractal Model of Internet Traffic*, Preprint, 2002.

[25] E. P. Tsolaki, *Non-stationary time series analysis of monthly global temperature data*, unpublished PhD thesis, University of Manchester Institute of Science and Technology (UMIST), 2001.

[26] J. Yuan, *Testing Gaussianity and linearity for random fields in the frequency domain.*, J. Time Ser. Anal., 21, no. 6, 723-737, 2000.

[27] J. Yuan, *Tests of Gaussianity and linearity for random fields using estimated higher order spectra*, IEEE Trans. on Signal Processing, 46, no. 1, 247-250, 1998.

[28] Y. Zhang, V. Paxson, S. Shenker, *The stationarity of Internet Path Properties: Routing, Loss, and Throughput*, ACIRI Technical Report, May 2000.