Signal Processing 89 (2009) 2537–2546

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Improved bispectrum based tests for Gaussianity and linearity

Yngve Birkelund *, Alfred Hanssen

Department of Physics and Technology, University of Tromsø, NO-9037 Tromso, Norway

ARTICLE INFO

Article history:
Received 22 February 2008
Received in revised form
7 April 2009
Accepted 10 April 2009
Available online 23 April 2009

Keywords:
Time series analysis
Gaussianity
Linearity
Bispectrum
Hypothesis tests
Surrogate data

ABSTRACT

The classical bispectrum based tests for linearity of time series are based on Gaussian asymptotics and a suboptimal smoothing in the bispectral domain. We show that the resulting classical tests may lead to vastly incorrect significance levels for non-Gaussian time series. This implies that a non-Gaussian linear time series may incorrectly be classified as non-linear. The purpose of this paper is to propose simple yet accurate tests for Gaussianity and linearity. The improved tests are derived through: (1) an optimal hexagonal smoothing in the bispectral domain, (2) the construction of simple and intuitive bispectrum based test statistics, and (3) determination of correct significance levels through a new skewness preserving scheme for linear surrogate data. The superiority of the proposed tests is demonstrated through extensive Monte Carlo simulations using relevant synthetic data.

1. Introduction

The linear Gaussian model has dominated time series analysis and modeling for decades [1–3]. However, the analysis of real-world data often reveals an underlying non-Gaussian and/or non-linear structure of the time series under scrutiny [4–6]. When dealing with time series data of unknown structure, a natural first step is to classify the data as accurately as possible. Thereafter, modeling, detailed analysis and interpretation may take place. The aim of the present paper is to suggest and demonstrate accurate and useful tests based on carefully constructed surrogate data, to aid in the classification of non-trivial time series data.

The classical statistical tests for non-Gaussianity and non-linearity [7,8] are based on the bispectrum [9] of the time series. Various classification statistics were compared in [10]. Alternative tests for non-linearity are, e.g., Ramsey's mis-specification test [11], White's neural network test [12], Paluš' information-theoretic redundancy approach [13] and the correlation dimension approach by Brock et al. [14]. Several tests were reviewed and compared in [15], and in [16], the authors suggested the use of bootstrapped residuals for Hinich's classical asymptotic test.

In this paper, we will concentrate on the classical tests by Hinich [8]. We will briefly review the estimation of the required power and bispectral densities, and point out certain severe statistical problems with the classical tests in their original form. In particular, Hinich's linearity test [8] does not provide the correct false alarm rate, and the suggested improvement in [16] does not necessarily correct the false alarm rate. Instead, it introduces further problems since the null hypothesis of the Gaussianity test is not fulfilled.

In the approach presented in this paper, we will advocate the use of carefully designed surrogate data to improve the classical tests so that their basic structure remains unchanged, while their statistical performance approaches their theoretical limit. The correct significance levels and thresholds can readily be found by means of the proposed surrogate data generator, also for the classical tests. Finally, we propose deceptively simple and intuitive test statistics for Gaussianity and linearity based on the skewness function. The proposed tests outperform the

* Corresponding author. Fax: +47 77 64 55 80.
E-mail addresses: yngve.Birkelund@phys.uit.no, yngveb@gmail.com (Y. Birkelund), alfred@phys.uit.no (A. Hanssen).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.04.013

classical tests in our extensive Monte Carlo simulations, and it clearly demonstrates the potential of the method of surrogates for practical data analysis.

The paper is organized as follows. The skewness function is defined and discussed in Section 2 as the central quantity for bispectrum based tests for non-Gaussianity and non-linearity. In Section 3 we review the classical asymptotical statistical hypothesis tests and reveal their problems. In Section 4 we present improved estimators for the skewness function and introduce non-Gaussian linear surrogate data for non-linear hypothesis tests. Original and improved tests are compared with numerical example in Section 5, before we summarize our findings and conclusions in Section 6.

2. The skewness function

2.1. Definitions

Assume the existence of a stationary real-valued discrete time zero-mean random process $\tilde{x}[n]$, for $n \in \mathbb{Z}$. The variance and skewness coefficient of this time series is defined as $\sigma^2 = E\{\tilde{x}^2[n]\}$ and $\gamma = E\{\tilde{x}^3[n]\}/(\sigma^2)^{3/2}$, respectively, where $E\{\cdot\}$ denotes the expectation operator.

The basic quantity arising in the classical tests for Gaussianity and linearity is the complex valued skewness function [9]

$$\Gamma(f_1,f_2) = \frac{S_3(f_1,f_2)}{\sqrt{S_2(f_1)S_2(f_2)S_2(f_1+f_2)}}, \tag{1}$$

where the power spectrum $S_2(f)$ is defined by

$$S_2(f) = \sum_{\tau=-\infty}^{\infty} R_2[\tau]\exp(-j2\pi f\tau), \tag{2}$$

where $R_2[\tau] = E\{\tilde{x}[n]\tilde{x}[n-\tau]\}$, and the bispectrum $S_3(f_1,f_2)$ is defined by

$$S_3(f_1,f_2) = \sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty} R_3[\tau_1,\tau_2]\exp[-j2\pi(f_1\tau_1+f_2\tau_2)], \tag{3}$$

where $R_3[\tau_1,\tau_2] = E\{\tilde{x}[n]\tilde{x}[n-\tau_1]\tilde{x}[n-\tau_2]\}$. For bispectrum based tests for Gaussianity and linearity the bifrequency region of interest is known as the *principal domain* (PD) of bispectral estimation [7,8,17], defined by the triangular region defined by the two inequalities $0 \leqslant f_1 \leqslant f_2 \leqslant 1/2$ and $2f_1 + f_2 \leqslant 1$.

Based on the skewness function, one can construct hypothesis tests for Gaussianity and linearity [7,8]. The reason is that theoretically, the skewness function of a Gaussian time series is identically zero everywhere, while the skewness function of a non-Gaussian linear times series has a constant non-zero magnitude [9]. A non-linear time series, on the other hand, exhibits a skewness function with bifrequency dependent magnitude. Since a Gaussian time series has a zero valued skewness function, we first have to test for Gaussianity. If the time series is found to be non-Gaussian, we can proceed to a second test to decide whether the time series follows a non-Gaussian linear model.

2.2. Estimators

In practice, only a finite length portion of a single realization of the process may be available, $x[n]; n = 0,1,\ldots,N-1$. Obviously, one will have to deal with estimates of the power spectrum $S_2(f)$ and the bispectrum $S_3(f_1,f_2)$. Hence, the sample version of the skewness function takes the form

$$\widehat{\Gamma}[k,l] = \frac{\widehat{S}_3[k,l]}{\sqrt{\widehat{S}_2[k]\widehat{S}_2[l]\widehat{S}_2[k+l]}}, \tag{4}$$

where $\widehat{S}_2[k]$ is a power spectrum estimate, $\widehat{S}_3[k,l]$ is a bispectrum estimate and k and l are discrete frequency indices. The statistical properties of the estimated skewness function come directly into play when choosing thresholds and confidence levels for the Gaussianity and linearity tests.

Before we enter a discussion about the skewness function as a test statistic, we will briefly review the class of spectral and bispectral estimators most commonly encountered. We will assume weak ergodicity, stationarity up to the third order, unity sampling interval, and that the power and bispectrum is to be estimated from a single realization $x[n]$ of length N of the time series. In the following, $X[k] = \sum_{n=0}^{N-1} x[n]\exp(-j2\pi kn/N)$ denotes the discrete Fourier transform (DFT) of $x[n]$, where the index $k = 0,1,\ldots,N-1$ corresponds to the discrete frequencies k/N.

2.2.1. Frequency smoothing

Smoothing in the frequency and the bifrequency domains is a simple and useful way to attain control over the estimator variance. Our basic frequency smoothed spectral and bispectral estimators are defined by

$$\widehat{S}_2[k] = \frac{1}{N}\sum_{k'=k-a}^{k+a} W_2[k']|X[k']|^2 \tag{5}$$

$$\widehat{S}_3[k,l] = \frac{1}{N}\sum_{k'=k-a}^{k+a}\sum_{l'=l-a}^{l+a} W_3[k',l']X[k']X[l']X^*[k'+l'] \tag{6}$$

respectively, where $W_2[k]$ and $W_3[k,l]$ are uniform smoothing windows and a is a smoothing bandwidth parameter. The statistical properties of the power and bispectral estimators depend strongly on the details of the chosen smoothing windows, and on the chosen smoothing bandwidth. For slowly varying power and bispectrum, unbiased estimates can be obtained for properly normalized smoothing windows. It is important to note that the ideal shape of the bispectral smoothing window $W_3[k,l]$ has a hexagonal region of support due to the fundamental symmetry properties of the bispectrum [18,19].

If we choose not to smooth at all, i.e., $a = 0$, the estimators in Eq. (5) and (6) reduce to the well-known periodogram and biperiodogram [20]. If we assume that the time series is Gaussian, the periodogram and biperiodogram are asymptotically unbiased, independent and distributed as chi-square and complex Gaussian random variables, respectively. Furthermore, the asymptotic variances of the periodogram and biperiodogram