

Testing for nonlinear serial dependence in time series with surrogate data and entropy measures

Test per la dipendenza seriale non lineare in serie storiche tramite surrogati e misure di entropia

Simone Giannerini, Esfandiar Maasoumi and Estela Bee Dagum

Abstract In this work we propose a nonparametric test for the identification of nonlinear dependence in time series. The approach is based on a combination of a test statistic based on an entropy dependence metric together with a suitable extension of surrogate data methods, a class of Monte Carlo tests introduced in the field of nonlinear dynamics. We focus on the null hypothesis of linear Gaussian processes and we derive the asymptotic theory for the test statistics. Since the asymptotic approximations depend on unknown quantities and require long series to be feasible we advocate the use of surrogate methods. We prove the asymptotic validity of the inference derived from the test and show the finite sample performance through a small simulation study.

Abstract In questo lavoro si introduce un test non parametrico per la dipendenza seriale non lineare in serie storiche. L'approccio è basato su di una misura metrica di entropia combinata con un estensione del metodo dei surrogati, una classe di test Monte Carlo introdotti nell'ambito della teoria dei sistemi dinamici non lineari. In questa sede si studia l'ipotesi nulla che il processo osservato sia lineare e gaussiano. A tal proposito, si ricava la distribuzione asintotica della statistica test sotto H_0 ; l'approssimazione asintotica dipende da quantità non note e richiede serie molto lunghe per poter essere applicata. Per tali motivi si utilizza il metodo dei surrogati e si dimostra la validità asintotica dell'inferenza ad esso associata. La performance del test su campioni finiti è studiata per mezzo di un piccolo studio di simulazione.

Simone Giannerini

Dipartimento di Scienze Statistiche, Università di Bologna, Via Belle Arti 41, 40126, Bologna, Italy. e-mail: simone.giannerini@unibo.it

Esfandiar Maasoumi

Department of Economics, Emory University, 1602 Fishburne Dr, Economics, Atlanta, Georgia, 30322, USA. e-mail: esfandiar.maasoumi@emory.edu

Estela Bee Dagum

Dipartimento di Scienze Statistiche, Università di Bologna, Via Belle Arti 41, 40126, Bologna, Italy. e-mail: estela.beedagum@unibo.it

Key words: Nonlinearity, Time series, Test, Entropy, Surrogate data

1 Introduction

In time series analysis it is important to ascertain whether the process under scrutiny cannot be adequately modelled by resorting to the linear gaussian paradigm. A practical difficulty that arise in such exercise is the apparent asymmetry between the realm of linearity and that of nonlinearity. In fact, while the class of linear processes is well defined from a mathematical point of view, the same does not hold for nonlinear processes. Thus, even though departures from linearity can occur in many directions, testing for nonlinearity is often a test for a specific nonlinear feature such as initial value sensitivity, fractality, volatility, threshold effects, irreversibility and so on. In other instances, nonlinearity is inferred from the failure of linear modelling and the problem reduces to either a diagnostic test (usually performed on the residuals of a linear model) or a specification test between models. For a recent review on the topic see [3] and references therein.

In this paper we study a non-parametric tests based on surrogate data together with an entropy measure of dependence. Our null hypothesis is motivated by the formal definition of linear processes as follows

$$H_0 : X_t = \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t \text{ with } \{\varepsilon_t\} \sim \text{i.i.d. } N(0, \sigma_\varepsilon^2), \quad (1)$$

where $\sum_{j=1}^{\infty} \phi_j^2 < \infty$ and $E[X_t]^4 < \infty$. In practice, H_0 assumes that the data generating process $\{X_t\}$ is a zero mean linear Gaussian stationary process. The alternative hypothesis H_1 states that $\{X_t\}$ does not admit a linear representation as in (1). Given a time series $\mathbf{x} = (x_1, \dots, x_n)$ and we are going to test whether $\mathbf{x}$ might be operationally considered as a realization of the process of Eq. (1).

2 The test statistic

The test statistic we propose is based upon the metric entropy measure S_ρ , a normalized version of the Bhattacharya-Hellinger-Matusita distance:

$$S_\rho(k) = \frac{1}{2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[[f_{(X_t, X_{t+k})}(x_1, x_2)]^{1/2} - [f_{X_t}(x_1) f_{X_{t+k}}(x_2)]^{1/2} \right]^2 dx_1 dx_2 \quad (2)$$

where $f_{X_t}(\cdot)$ and $f_{(X_t, X_{t+k})}(\cdot, \cdot)$ denote the probability density function of X_t and of the vector (X_t, X_{t+k}) respectively. The good theoretical properties of the measure are discussed in [5, 9, 10]. In particular, for the construction of a test statistic we use the

relationship of S_ρ with the autocorrelation coefficient in the Gaussian case. We find the following:

Proposition 1. *Let $(X, Y) \sim N(\mathbf{0}, \mathbf{1}, \rho)$ be a standard Normal random vector with joint probability density function given by $f_{X,Y}(\cdot, \cdot, \rho)$, where ρ is the correlation coefficient. Then, the following relation holds:*

$$S_\rho = 1 - \frac{2(1 - \rho^2)^{1/4}}{(4 - \rho^2)^{1/2}} \quad (3)$$

Proof. See [4] $\square$

For the sake of brevity, in the following, instead of $S_\rho(k)$ we will use S_k . Now, by using the result of Eq. (3) we can obtain an estimator for S_k under the null of a linear Gaussian process based on the sample autocorrelation ρ_k . We denote such a parametric estimator with $\hat{S}_k^p$ where the superscript p stands for *parametric*. The consistency of the estimator $\hat{S}_k^p$ and asymptotic normality under H_0 have been derived in [4]. In order to test the null hypotheses of linearity H_0 (1) we propose the following test statistic:

$$\hat{T}_k = [\hat{S}_k^u - \hat{S}_k^p]^2 \quad (4)$$

where $\hat{S}_k^u$ is the nonparametric estimator for S_k based on kernel density estimators and numerical quadrature for the integral of Eq (2); here we have adopted the maximum likelihood cross validation as bandwidth selection method. In other words, $\hat{T}_k$ is the squared divergence between the unrestricted nonparametric estimator and the parametric estimator of S_k under H_0 . The following theorem establishes strong convergence and the asymptotic distribution of $\hat{T}_k$ under the null.

Theorem 1. *Under H_0*

1. $\hat{T}_k \rightarrow 0 \quad \text{in } L^2.$
2. $\frac{n\hat{T}_k}{\sigma_a^2} \rightarrow \chi_1^2 \quad \text{in distribution.}$

where σ_a^2 is the asymptotic variance of $\hat{T}_k^{1/2}$.

Proof. See [4]

Theorem 1 assures that the test statistic will converge to zero if the process is linear and Gaussian. Hence, large values of $\hat{T}_k$ will indicate departure from the hypothesis of linearity. The derivation of the asymptotic approximation for the significance level and power of the test is bound to deriving an estimator for the asymptotic variance σ_a^2 . The task is practically unfeasible and the exercise is of limited practical relevance. Moreover, several contributions in literature show that very large sample sizes are required to obtain meaningful results (see e.g. [6, 7, 15, 8]).

In order to show the performance of the asymptotic χ_1^2 approximation of the test based on $\hat{T}_k$. We generate 500 series with $n = 2000$ from the moving average process $X_t = 0.8\varepsilon_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim N(0, 1)$ for all t . Recall that $\rho_1 = -0.488$ and $\sigma_a^2 = \sigma_p^2 + \sigma_u^2$. Now, σ_p^2 is the asymptotic variance of the parametric estimator $\hat{S}_1^p$ and can be computed exactly for such process (see [4]):

$$\sigma_p^2 = \left[\frac{\rho_1(2 + \rho_1^2)}{(1 - \rho_1^2)^{3/4}(4 - \rho_1^2)^{3/2}} \right]^2 (1 + 2\rho_1^2) = 0.04966.$$

As for σ_u^2 , we derive an estimate for it through the Monte Carlo variance of $\hat{S}_1^u$. The histogram of $\frac{n\hat{T}_1}{\sigma_a^2}$ is shown in Fig. 1 where we have superimposed a χ_1^2 density (solid red line). We are in the unrealistic scenario in which we know exactly the asymptotic variance of the parametric estimator and have a very good estimate for that of the non-parametric estimator; still, the approximation of the right tail of the distribution of the test statistic is not satisfactory. In fact, if we set the nominal level $\alpha = 0.05$ and use the χ_1^2 approximation we get an empirical rejection rate of 0.09. For these

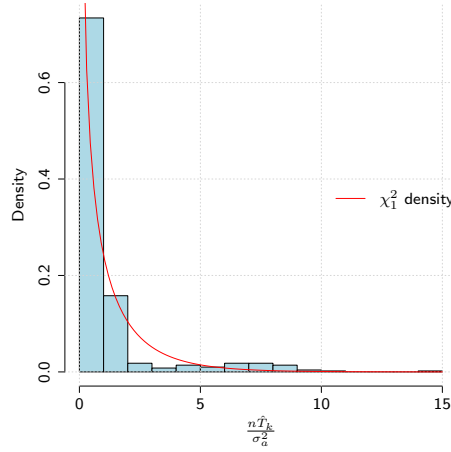


Fig. 1 Histogram of $\frac{n\hat{T}_1}{\sigma_a^2}$ for 500 realizations of a moving average process of order 1. The length of the series is $n = 2000$. The solid red line is a χ_1^2 density.

reasons, we propose a resampling scheme based on the method of surrogate data that leads to valid inference and good finite sample performance.

3 The surrogate data approach

The method of surrogate data first appeared in the literature of nonlinear dynamics and chaos theory (see [13]). Loosely speaking, it can be seen as a nonparamet-

ric resampling approach aimed at building tests for nonlinearity. The key of the method lies in generating surrogate series having the same periodogram and the same marginal distribution as the original series. This is achieved by randomizing the phase of the sample periodogram of the series, for more details see [13] and [14]. In [1] it is proved that under the null hypothesis that the generating process is a stationary Gaussian circular process such method produces exactly valid inferences. In other words, the resulting tests have a Neyman structure. The asymptotic validity for the null hypothesis of a stationary Gaussian process is also given (see also [2], Chap. 4.4). For further discussions and some extensions of the method see [12].

In this paper we produce surrogate series possessing the aforementioned properties by using a constrained randomization scheme (see also [11, 12]). Rather than focusing on the frequency domain the procedure generates surrogate time series having the same autocorrelation structure and marginal distribution of the original series by solving a stochastic optimization problem through simulated annealing. In practice: *i*) Define one or more constraints in terms of a cost function that reaches a global minimum when the constraints are fulfilled; *ii*) Minimize the cost function among all the possible permutations of the series through simulated annealing. In turn, the cost function is defined in terms of the distance between the sample autocorrelation of the original and that of the surrogate series. In [4] we have proved that that under the null H_0 the surrogate approach combined with the test statistic T_k leads to asymptotically valid inferences.

4 Some results

In this section we assess the performance of the tests in finite samples by means of a small simulation study. We focus on the following models:

Model 1: AR(1) $x_t = 0.8x_{t-1} + \varepsilon_t$

Model 2: ARt(1) $x_t = 0.8x_{t-1} + \zeta_t$

Model 3: NLMA $x_t = 0.8\varepsilon_{t-2}^2 + \varepsilon_t$

Model 4: SETAR $x_t = \begin{cases} -0.8x_{t-1} + \varepsilon_t & \text{if } x_{t-1} \leq 0 \\ 0.8x_{t-1} + \varepsilon_t & \text{if } x_{t-1} > 0 \end{cases}$

where the innovation processes are i.i.d. $\varepsilon_t \sim N(0,1)$ and $\zeta_t \sim \text{Student's } t$ with 3 degrees of freedom. Model 1 is a linear Gaussian processes whereas Model 2 is linear with non-Gaussian innovations. Model 3 is a nonlinear moving average process and Model 4 is a threshold autoregressive process. For each process we have generated 500 series with $n = 200$. In Figures 2 and 3 we show the mean over 500 Monte Carlo replications of the test statistic $\hat{T}_k$ (black solid line) together with the rejection bands at level 95% and 99% (green and blue dashed lines), derived from the quantiles of the distributions of the surrogates. The abscissa indicates the lag $k = 1, \dots, 5$.

Figure 2 shows that the test manages to distinguish between linear Gaussian and linear non-Gaussian processes. Moreover, in Figure 3 it identifies the lags at which

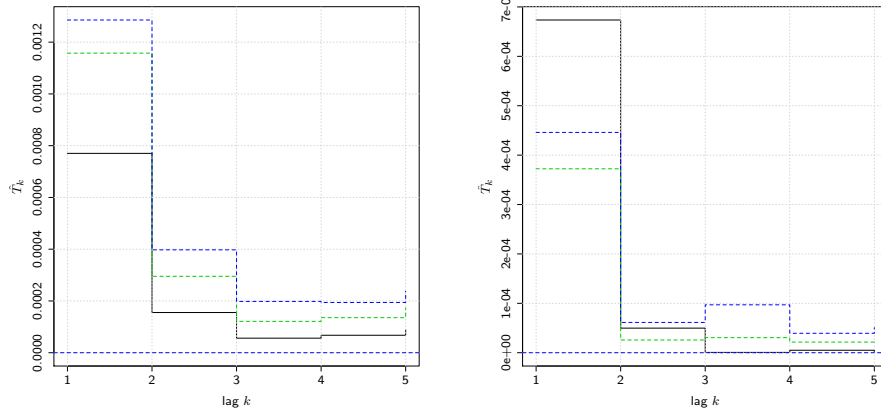


Fig. 2 (Left): mean $\hat{T}_k$, $k = 1, \dots, 5$ for 500 realizations of Model 1 (Gaussian autoregressive process, black solid line). The rejection bands at 95% and 99% level (green/light gray and blue/dark gray dashed line, respectively) are obtained from the mean quantiles of the surrogates distribution. (Right) the same but for Model 2 (linear, non-Gaussian process).

the different kinds of nonlinearity implied by models 3 and 4 are supposed to manifest. Contrarily to the majority of existing proposal, the test does not need assumptions on the nature of the data generating process. Also, it manages to identify the lags at which a nonlinear effect is expected. For a more detailed analysis on power, size, computational complexity, etc. see [4]. Further investigations will include studying the impact of different bandwidth selection methods. The test has been implemented in the R package `tseriesEntropy` which is forthcoming on CRAN (see www2.stat.unibo.it/giannerini/software.html).

Acknowledgements Simone Giannerini acknowledges the support of the PRIN project 2010-2011, (prot. 2010J3LZEN): La previsione economica e finanziaria: il ruolo dell'informazione e la capacità di modellare il cambiamento.

References

1. Chan, K.: On the validity of the method of surrogate data. In: C.D. Cutler, D.T. Kaplan (eds.) *Nonlinear Dynamics and Time Series*, *Fields Inst. Communications*, vol. 11, pp. 77–97. American Math. Soc. (1997)
2. Chan, K., Tong, H.: *Chaos: a statistical perspective*. Springer Verlag, New York (2001)
3. Giannerini, S.: The quest for nonlinearity in time series. In: T. Subba Rao, C. Rao (eds.) *Handbook of Statistics: Time Series*, *Handbook of Statistics*, vol. 30, pp. 43–63. Elsevier (2012)
4. Giannerini, S., Maasoumi, E., Bee Dagum, E.: Entropy testing for nonlinear serial dependence in time series. Under review (2014)
5. Granger, C., Maasoumi, E., Racine, J.: A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis* **25**(5), 649–669 (2004)
6. Hjellvik, V., Tjøstheim, D.: Nonparametric tests of linearity for time series. *Biometrika* **82**(2), 351–368 (1995)

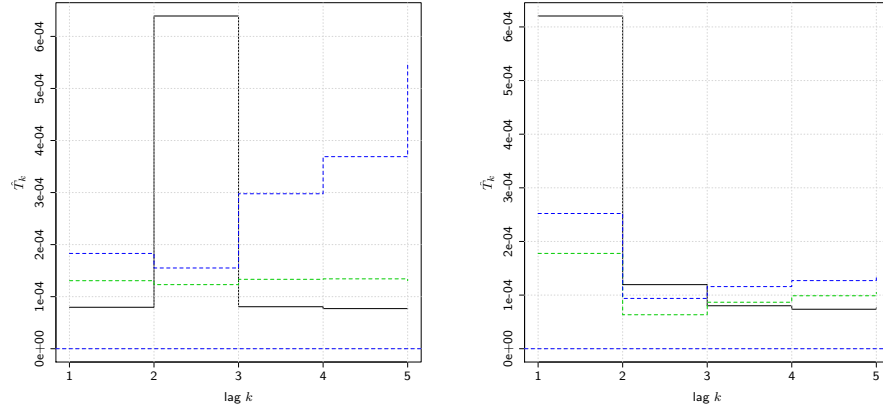


Fig. 3 (Left): mean $\hat{T}_k$, $k = 1, \dots, 5$ for 500 realizations of Model 3 (nonlinear moving average process, black solid line). The rejection bands at 95% and 99% level (green/light gray and blue/dark gray dashed line, respectively) are obtained from the mean quantiles of the surrogates distribution. (Right) the same but for Model 4 (threshold autoregressive process).

7. Hjellvik, V., Yao, Q., Tjøstheim, D.: Linearity testing using local polynomial approximation. *Journal of Statistical Planning and Inference* **68**(2), 295 – 321 (1998)
8. Hong, Y., White, H.: Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* **73**(3), 837–901 (2005)
9. Maasoumi, E.: A compendium to information theory in economics and econometrics. *Econometric Rev.* **12**(2), 137–181 (1993)
10. Micheas, A., Zografos, K.: Measuring stochastic dependence using φ -divergence. *Journal of Multivariate Analysis* **97**(3), 765 – 784 (2006)
11. Schreiber, T.: Constrained randomization of time series data. *Phys. Rev. Lett.* **90**(10), 2105–2108 (1998)
12. Schreiber, T., Schmitz, A.: Surrogate time series. *Physica D* **142**(3–4), 346–382 (2000)
13. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.: Testing for nonlinearity in time series: the method of surrogate data. *Physica D* **58**, 77–94 (1992)
14. Theiler, J., Prichard, D.: Constrained-realization monte-carlo method for hypothesis testing. *Physica D* **94**, 221–235 (1996)
15. Tjøstheim, D.: Measures of dependence and tests of independence. *Statistics: A Journal of Theoretical and Applied Statistics* **28**(3), 249–284 (1996)