Detecting Non Linearity in Time Series by Model Selection Criteria

by

Daniel Peña[1] and Julio Rodriguez[2]

[1]Departamento de Estadística. Universidad Carlos III de Madrid. Getafe 28903. Spain.

E-mail: daniel.pena@uc3m.es

[2]Laboratorio de Estadística, ETSII, Universidad Politécnica de Madrid. Jose Gutierrez Abascal, 2, 28006 Madrid.

E-mail: puerta@estsii.upm.es

Brief Title: **Detecting Non Linearity**

Abstract

This article analyses the use of model selection criteria for detecting non linearity in the residuals of a linear model. Model selection criteria are applied for finding the order of the best autorregressive model fitted to the squared residuals of the linear model. If the order selected is not zero, this is considered as an indication of non linear behavior. The BIC and AIC criteria are compared in three Monte Carlo experiments to some popular nonlinearity tests. We conclude that the BIC model selection criterion seems to offer a promising tool for detecting non linearity in time series. An example is shown to illustrate the performance of the tests considered and the relationship between non linearity and structural changes in time series.

Key Words and Phrases: AIC, BIC, Bilinear, GARCH, Portmanteau tests, Threshold autoregressive.

AMS 1991 Subject Classifications. Primary: 62M10; **Secondary:** 62M20.

1 Introduction

Non linear time series models have received a growing interest both from the theoretical and the applied points of view. See the books by Priestley (1989), Tong (1990), Granger and Teräsvirta (1992), Terdik (1999), Peña, Tiao and Tsay (2001) and Fan and Yao (2003), among others. Non linearity testing has been an active subject of research. First, some tests were developed based on the frequency domain approach by using the biespectral density function. See Subba Rao and Gabr (1980) and Hinich (1982), among others. Second, the Volterra expansion suggests testing for non linearity by using the residuals of the linear fit and by introducing added variables which can capture nonlinear effects. Keenan (1985), Tsay (1986, 1991, 2001) and Luukkonen et al. (1988), among others, have proposed specific tests based on this idea. Third, we can use a non parametric approach as in Hjellvik and Tjϕstheim (1995) and Brock, Dechert, Scheinkman and LeBaron (1996).

Another way to obtain a non linear test is by noting that if the residuals of the linear fit $\hat{\varepsilon}_t$ are not independent, they could be written as

$$\hat{\varepsilon}_t = m(\hat{\boldsymbol{\varepsilon}}_{t-1}) + u_t v(\hat{\boldsymbol{\varepsilon}}_{t-1}), \tag{1}$$

where $\hat{\boldsymbol{\varepsilon}}_{t-1} = (\hat{\varepsilon}_{t-1}, ..., \hat{\varepsilon}_1)$ is the vector of past residuals and u_t is a sequence of zero mean and unit variance i.i.d random variables independent of $\hat{\boldsymbol{\varepsilon}}_{t-1}$. Assuming that the residuals follow a zero mean stationary sequence, we have that $E(m(\hat{\boldsymbol{\varepsilon}}_{t-1})) = 0$ and

$E(m^2(\hat{\boldsymbol{\varepsilon}}_{t-1})) = c_1$, and $E(v^2(\hat{\boldsymbol{\varepsilon}}_{t-1})) = c_2$, where c_1 and c_2 are constants. Expression (1) includes, among others, bilinear, TAR (threshold autoregressive), STAR (smooth transition threshold autoregressive) ARCH (autoregressive conditional heteroskedastic) and GARCH models. Making a Taylor expansion around the zero mean value of the residuals in (1), and by using that the residuals should be uncorrelated, we have

$$\hat{\varepsilon}_t = \frac{1}{2}\sum\sum \frac{\partial^2 m(\hat{\boldsymbol{\varepsilon}}_{t-1})}{\partial\hat{\varepsilon}_{t-j}\partial\hat{\varepsilon}_{t-g}}\hat{\varepsilon}_{t-j}\hat{\varepsilon}_{t-g} + u_t\left[\sum \frac{\partial v(\hat{\boldsymbol{\varepsilon}}_{t-1})}{\partial\hat{\varepsilon}_{t-j}}\hat{\varepsilon}_{t-j} + \frac{1}{2}\sum\sum \frac{\partial^2 v(\hat{\boldsymbol{\varepsilon}}_{t-1})}{\partial\hat{\varepsilon}_{t-j}\partial\hat{\varepsilon}_{t-g}}\hat{\varepsilon}_{t-j}\hat{\varepsilon}_{t-g}\right] + R$$

where R includes higher order terms and it is of order smaller than $1/T$. Taking the squared of this expression and computing the conditional expectation given the past values then, approximately, we can write

$$E(\hat{\varepsilon}_t^2|\hat{\boldsymbol{\varepsilon}}_{t-1}) = c + \sum a_i\hat{\varepsilon}_{t-i}^2 + \sum\sum b_{ij}\hat{\varepsilon}_{t-i}^2\hat{\varepsilon}_{t-j}^2 + R' \qquad (2)$$

where c is a constant and R' includes terms of order equal or higher than 3. This equation implies a complex autoregressive dependency among the squared residuals, and suggests that we can test for non linear behavior by analyzing the presence of linear dependency among the squared residuals. This idea was proposed by Granger and Andersen (1978) and Maravall (1983), and it has been used for McLeod and Li (1983) and by Peña and Rodriguez (2002, 2005) for building portmanteau test of goodness of fit.

Finally, specific tests for a particular kind of non linearity have also been developed. In particular, these tests can be useful when the Taylor expansion which justifies (2) is not appropriate and therefore the power of global non linear tests based on squared residuals is expected to be low. For instance, in threshold autorregressive models (TAR) the non linear function which relates the series to its past is not smooth, and therefore we will need many terms in the Taylor expansion to approximate it. Tsay (1989) has developed a powerful test for checking for TAR behavior. Other area in which several specific tests have been proposed is conditional heteroskedastic models, see, among others, the Engle LM test to detect ARCH disturbances, the Harvey and Streibel (1998) test and the Rodriguez and Ruiz (2005) test. However, for these models procedures based on the squared residuals are expected to work well.

In this article we consider an alternative way to check if the residuals of a linear fit have linear dependency. Based on expression (2), we explore the performance of using a model selection criterion to obtain the order of the best autoregressive model fitted to the squared residuals. If the selected order is zero, we conclude that there is no indication of non linearity, whereas if the selected model is AR(p), $p > 0$, we conclude that the time series is non linear. A similar idea was advocated in the linear case by Pukkila et al. (1990). They proposed an iterative procedure for determining the order of ARMA(p,q) models which consists of fitting an increasing order ARMA structure to the data and verifying that the residuals behave like white noise by using an autoregressive order determination information criterion. They found that the BIC criterion worked very well in the linear case; see also Koreisha and Pukkila

(1995).

The rest of the paper is organized as follows. In section 2, we briefly describe the global non linearity tests that we will consider in the Monte Carlo analysis. In section 3, we discuss model selection criteria which can be used for fitting autoregressive models to the squared residuals of the linear fit. Section 4 presents the Monte Carlo study. Section 5 contains an example and Section 6 some concluding remarks.

2 Types of non-linear test

In this section we describe briefly four types of global non-linear tests which we will include in all the experiments of the Monte Carlo study. These four tests have been chosen by using two criteria. First, they are based on different principles and, second, all of them have shown a good performance for some class of non linear models in previous Monte Carlo experiments; see Tsay (1991), Lee, White and Granger (1993) and Ashley and Patterson (1998). The first two tests are based on the residuals of the linear fit. The Tsay test checks for the inclusion of added variables to represent the non linear behavior, whereas the BDS test relies on smoothness properties. The second two tests are based on the squared residuals. The McLeod and Li (1983) test uses the asymptotic sample distribution of the estimated autocorrelations, whereas the Peña and Rodriguez (2005) test uses the determinant of their correlation matrix.

Keenan (1985) proposed a test in which the residuals of a linear fit are related to a proxy variable of the non linear behavior in the time series, as follows: (1) A linear model is fitted to the time series $\widehat{y}_t = \sum_{i=1}^{M} \widehat{\alpha}_i y_{t-i}$, and the residuals of the linear fit, $\hat{\varepsilon}_t = y_t - \widehat{y}_t$, which will be free from linear effects, are computed; (2) A proxy variable for the non-linear part in the time series is obtained by $x_t = \widehat{y}_t^2 - \widehat{y}_t$; (3) A regression is made between these two variables, $\hat{\varepsilon}_t = \delta x_t + u_t$, and the non linearity test is the standard regression test for $\delta = 0$. Note that this test uses a proxy variable which includes jointly the squares and cross products of the M lags of the series.

The first test we will include in our Monte Carlo study is due to Tsay (1986), who generalizes the previous proposal by Keenan. Tsay improves this test by decomposing the proxy variable x_t into different regressors in a multiple linear regression equation. Thus, instead of using jointly the squared and cross product effects of the variables $(y_{t-1}, ..., y_{t-M})$ in $\widehat{y}_t^2$, $h = M(M+1)/2$ variables are defined which include all the possible squares and cross product terms of these lag variables. The test is implemented as follows: (1) Fit the linear model $\widehat{y}_t = \sum_{i=1}^{M} \widehat{\alpha}_i y_{t-i}$, and compute the residuals $\hat{\varepsilon}_t = y_t - \widehat{y}_t$; (2) Define $z_{1t} = y_{t-1}^2$, $z_{2t} = y_{t-1}y_{t-2}, ..., z_{Mt} = y_{t-1}y_{t-M}$, $z_{M+1,t} = y_{t-2}^2, z_{M+2,t} = y_{t-2}y_{t-3}, ..., z_{ht} = y_{t-M}^2$. Then, regress each of these h variables z_{jt} against $(y_{t-1}, ..., y_{t-M})$, and obtain the residuals, $x_{jt} = z_{jt} - \sum_{i=1}^{M} \widehat{\beta}_i^j y_{t-i}$, which will be our proxy variables for non linear behavior; (3) Regress $\hat{\varepsilon}_t$ to the h proxy variables x_{jt}, and compute the usual F statistic for testing that all the regression coefficients in the population are equal to zero. The linearity is rejected if the F test finds any proxy variable as significant to explain the residuals of the linear fit. Thus, in practice the null hypotheses of this test is that there is no linear relationship between the residuals of the linear fit and the set of proxy variables which include the squared and cross products terms. Note that the only parameter that needs to be defined in this test is the number of lags, M, used in the AR fitting. This test has been extended to include some specific forms of nonlinear models, see Tsay (1991, 2001) but in this paper we will use the original formulation.

The second test based on the residuals of the linear fit is the one by Brock et al. (1991, 1996). These authors proposed a test, called the BDS test in the literature, which has become quite popular. The idea of the test is as follows. No matter how the non linear relation is, if we start the time series by using the same starting values, the future values are expected to be similar, at least in the short run. Therefore, given two blocks of time series points

$$(\hat{\varepsilon}_t, ..., \hat{\varepsilon}_{t+k-1}) \quad \text{and} \quad (\hat{\varepsilon}_{t+s}, ..., \hat{\varepsilon}_{t+s+k-1}) \tag{3}$$

which are close in some metric, we expect that the future evolution of the next values after these two blocks, $(\hat{\varepsilon}_{t+k}, ..., \hat{\varepsilon}_{t+k+g})$ and $(\hat{\varepsilon}_{t+s+k}, ..., \hat{\varepsilon}_{t+s+k+g})$ should also be close in the same metric. These authors propose as measure of closeness the largest euclidean distance between members of the two blocks which have the same position. That is, if we consider the two sequences in (3), the distances $d_j = |\hat{\varepsilon}_{t+j} - \hat{\varepsilon}_{t+s+j}|$ for $j = 0, ..., k-1$ are computed and the sequences are judged to be close if $\max(d_j) \leq c$.

This idea is implemented in a test as follows. We form all possible sequences of k elements $(\hat{\varepsilon}_t, ..., \hat{\varepsilon}_{t+k-1})$, for $t = 1, ..., n-k$ and we count the number of other sequences of k consecutive elements which are close to the one we are considering. The result of comparing two sequences or blocks of size k, one starting at time s and the other starting at time t is given by a dummy variable, $C_{t,s}$, which takes the value one when the two sequences are close and zero otherwise. The comparison among all the sequences of size k is summarized by the proportion of them which are close, and this proportion is computed by

$$C_{k,T} = \frac{2}{(T-k)(T-k-1)} \sum_{t=1}^{T-k} \sum_{s=t+1}^{T-k-1} C_{t,s}.$$

The BDS test statistic is the standardized value of $C_{k,T}$:

$$w_{k,T} = \sqrt{T-k-1} \frac{(C_{k,T} - C_{1,T-k+1}^{k})}{\sigma_{k,T-k+1}}$$

which, under the hypotheses of independence, follows a normal distribution asymptotically. The null hypotheses of the test is that the number of sequences which are close in the residuals of the time series is similar to the number expected with independent data. In order to use this test we have to define k and d.

The third test we discuss is the one proposed by McLeod and Li (1983). They computed the squared residuals and their autocorrelations by

$$r_k = \frac{\sum_{t=k+1}^{T} \left(\hat{\varepsilon}_t^2 - \hat{\sigma}^2\right) \left(\hat{\varepsilon}_{t-k}^2 - \hat{\sigma}^2\right)}{\sum_{t=1}^{T} \left(\hat{\varepsilon}_t^2 - \hat{\sigma}^2\right)^2}, \quad (k = 1, 2, ..., m), \tag{4}$$

where $\hat{\sigma}^2 = \sum \hat{\varepsilon}_t^2/T$, and suggested checking for non linearity by using the Ljung-Box statistic but now applied to the autocorrelations among the squared residuals. The test statistic is

$$Q_{ML} = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}r_k^2. \tag{5}$$

They showed that, under the hypotheses of linearity, the statistic Q_{ML} follows aymptotically a χ_m^2 distribution. The null hypotheses of this test is that the first m autocorrelations among the squared residuals are zero. This test only depends on the parameter m. A similar test can be developed by using the statistic based on the partial autocorrelacions proposed by Monti (1994), but as its power is smaller than the next statistic (see Peña and Rodriguez, 2002, 2005), we have not included it in this study.

The fourth statistic we discuss here is the one proposed by Peña and Rodriguez (2005):

$$D_m = -\frac{T}{m+1}\log|\tilde{\mathbf{R}}_m|,$$

where $\tilde{\mathbf{R}}_m$ is the autocorrelation matrix of the standardized autocorrelation coefficients of the squared residuals $\tilde{r}_k$, defined by

$$\tilde{r}_k^2 = \frac{(T+2)}{(T-k)}r_k^2, \tag{6}$$

where r_k is given by (4), and

$$\widetilde{\mathbf{R}}_m = \begin{bmatrix} 1 & \widetilde{r}_1 & \cdots & \widetilde{r}_m \\ \widetilde{r}_1 & 1 & \cdots & \widetilde{r}_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{r}_m & \widetilde{r}_{m-1} & \cdots & 1 \end{bmatrix}. \tag{7}$$

Under linearity, this statistic follows aymptotically a gamma distribution, $\mathcal{G}(\alpha, \beta)$ where $\alpha = 3m(m+1)/4(2m+1)$ and $\beta = 3m/2(2m+1)$. The transformation $\mathrm{ND}_m = (D_m^\lambda - E(D_m^\lambda))/std(D_m^\lambda)$ where $\lambda = g(\alpha, \beta)$ is given in Peña and Rodriguez (2005) follows a standard normal variable. This test was obtained from multivariate analysis of covariance matrices and its null hypotheses is that the first m autocorrelations among the squared residuals are zero. These authors showed that this test is more powerful than other tests also based on the squared autocorrelations, including the one by McLeod and Li.

3 Model Selection Criteria

Suppose that we want to select the autoregressive order for a given time series. We cannot select the order by using the residual variance because this measure cannot increase if we increase the order of the autoregression. Similar problems appear with other measures of fit, as the deviance. Model selection criteria were introduced to solve this problem. The most often used criteria can be written as:

$$\min_k \left\{ \log \widehat{\sigma}_k^2 + k \times C(T,k) \right\}, \tag{8}$$

where $\widehat{\sigma}_k^2$ is the maximum likelihood estimate of the residual variance, k is the number of estimated parameters for the mean function of the process, T is the sample size and the function $C(T,k)$ converges to 0 when $T \to \infty$. These criteria have been derived from different points of view. Akaike (1969), a pioneer in this field, proposed selecting the model with the smallest expected out of sample forecast error, and derived an asymptotic estimate of this quantity. This led to the final prediction error criterion, FPE, where $C(T,k) = k^{-1}\log(\frac{T+k}{T-k})$. This criterion was further generalized, using information theory and Kullblack - Leibler distances, by Akaike (1973), in the well known AIC criterion where $C(T,k) = 2/T$. Shibata (1980) proved that this criterion is efficient, which means that if we consider models of increasing order with the sample size, the model selected by this criterion is the one which produces the least mean square prediction error. The AIC criterion has a bad performance in small samples because it tends to overparametrize too much. To avoid this problem, Hurvich and Tsai (1989) introduced the corrected Akaike's information criterion, $AICC$, where $C(T,k) = \frac{1}{k}\frac{2(k+1)}{T-(k+2)}$.

From the Bayesian point of view it is natural to choose among models by selecting the one with the largest posterior probability. Schwarz (1978) derived a large sample approximation to the posterior probability of the models assuming the same prior probabilities for all of them. The resulting model selection criterion is called the Bayesian information criterion, BIC, and in (8) it corresponds to $C(T,k) = \log(T)/T$. As the posterior probability of the true model will go to one when the sample size increase, it can be proved that BIC is a consistent criterion, that is, under the assumption that the data come from a finite order autoregressive moving average process, we have a probability of obtaining the true order that goes to one when $T \to \infty$. Other often used consistent criterion is the one due to Hannan and Quinn (1979), called HQC, where $C(T,k) = 2m \log\log(T)/T$ with $m > 1$.

Galeano and Peña (2004) proposed to look at model selection in time series as a discriminant analysis problem. We have a set of possible models, $M_1, ..., M_\alpha$, with prior probabilities $P(M_i)$, $\sum P(M_i) = 1$, and we want to classify a given time series, $\mathbf{y} = (y_1, ..., y_n)$ as generated from one of these models. The standard discriminant analysis solution to this problem is to classify the data in the model with highest posterior probability and, if the prior probabilities are equal, this leads to the BIC criterion. From the frequentist point of view the standard discriminant analysis solution when the parameters of the model are known is to assign the data to the model with the highest likelihood. If the parameters of the models are unknown we can estimate them by maximum likelihood, plug them in the likelihood, and select again the model with the highest estimated likelihood. However, although this procedure works well when we are comparing models with the same number of unknown parameters, it cannot be used when the number of parameters are different. As the estimated likelihood cannot decrease by using a more general model, the maximum estimated likelihood criterion will always select the model with more parameters. To avoid this problem, Galeano and Peña (2004) proposed to select the model which has the largest *expected* likelihood, as follows. Compute the expected value of the likelihood over all possible sequences generated by the model and choose the model with largest expected likelihood. These authors proved that the resulting procedure is equivalent to the AIC criterion.

4 Monte Carlo Experiments

In these section we present three experiments to compare non linearity tests and model selection criteria for detecting non linear behavior in time series. The first experiment is designed to compare the size and power of the methods under investigation when the process is non linear in the mean function, but has a constant variance. The second experiment compares them for detecting non linearity in the variance function, as in the ARCH, GARCH and SV processes. The third experiment replicates the design of the competition among non linear test in Barnett et al. (1997), which includes also deterministic chaos as well as non linearity, either in the mean or in the variance function.

In the first experiment ten non linear models were used. These models have been

M1:	$Y_t = 0.4Y_{t-1} + 0.8Y_{t-1}\varepsilon_{t-1} + \varepsilon_t$
M2:	$Y_t = \begin{cases} 1 - 0.5Y_{t-1} + \varepsilon_t & Y_{t-1} \leq 1 \\ 1 + 0.5Y_{t-1} + \varepsilon_t & Y_{t-1} > 1 \end{cases}$
M3	$Y_t = \begin{cases} 1 - 0.5Y_{t-1} + \varepsilon_t & Y_{t-1} \leq 1 \\ 1 + \varepsilon_t & Y_{t-1} > 1 \end{cases}$
M4:	$Y_t = -0.4\varepsilon_{t-1} + 0.3\varepsilon_{t-2} + 0.5\varepsilon_t\varepsilon_{t-2} + \varepsilon_t$
M5:	$Y_t = -0.3\varepsilon_{t-1} + 0.2\varepsilon_{t-2} + 0.4\varepsilon_{t-1}\varepsilon_{t-2} - 0.25\varepsilon_{t-2}^2 + \varepsilon_t$
M6:	$Y_t = 0.4Y_{t-1} - 0.3Y_{t-2} + 0.5Y_{t-1}\varepsilon_{t-1} + \varepsilon_t$
M7:	$Y_t = 0.4Y_{t-1} - 0.3Y_{t-2} + 0.5Y_{t-1}\varepsilon_{t-1} + 0.8\varepsilon_{t-1} + \varepsilon_t$
M8:	$Y_t = 0.2\varepsilon_{t-1}^3 + \varepsilon_t$
M9:	$Y_t = 0.6\varepsilon_{t-1}\left[\varepsilon_{t-2}^2 + 0.8\varepsilon_{t-3}^2 + 0.8^2\varepsilon_{t-4}^2 + 0.8^3\varepsilon_{t-5}^2\right] + \varepsilon_t$
M10a:	$Y_t = 0.5Y_{t-1} + \varepsilon_t$
M10b:	$Y_t = 0.3Y_{t-1} + 0.5Y_{t-2} - 0.5Y_{t-3} + \varepsilon_t$

Table 1: Models included in the first Monte Carlo experiment

previously proposed in the literature for comparing non linear tests and are presented in Table 1, were throughout $\varepsilon_t \sim N(0,1)$ is a white noise series. Models M1, M2 and M3 were analyzed by Harvill (1999), models M4, M5, M6 and M7 by Keenan (1985), and models M8 and M9 by Ashley and Patterson (1998). Models M10a and M10b are linear AR(1) and AR(3) models and they are used to compute the size of the tests.

The experiment was run as follows. For each model in Table 1 and one of the three sample sizes considered, $n = 50, 100, 250$, a time series was generated. A linear $AR(p)$ model was fitted to the data, where p was selected by the AIC criterion, (Akaike, 1974) with $p \in \{1, 2, 3, 4\}$, and the residuals were computed. Then, the four linearity tests described in the previous section were applied. The value of M in the test by Tsay is $M = 5$, and this test will be indicated in the tables by F^5_{Tsay}. In the BDS test the parameters are $k = 2, 3, 4$, and $d = \varepsilon/\sigma = 1.5$; the corresponding results of the test will be given in the tables under the heading BDS_2, BDS_3, BDS_4. The value of m for both the D_m and Q_{ML} tests is $m = [\sqrt{T}]$, and the maximum AR order is $p_{max} = [\sqrt{T}]$. The best model is then selected by the three criteria considered, AIC, AICC, BIC. For each model and sample size 5000 runs were made.

We first present in Table 2 the size of the tests when the time series is really generated by a linear model, M10a, or M10b. Columns 2 to 6 of this table show the proportion of the 5000 runs in which a given test rejects the hypotheses of linearity when the test is applied with a significant level of .95. Columns 7 to 9 show the proportions in which the model selection criterion selects a value greater than zero as best order for the squared residuals. The results in this table indicate that for $n = 250$ all the tests have sizes close to the nominal value, .05. The size of the tests D_m, Q_{ML} and F^5_{Tsay} improve with T and get close to the value .05 when the sample size increases. This is in agreement with the fact that we are using asymptotic critical percentiles. However, for the BDS test instead of the asymptotic percentiles we have used the estimated finite sample empirical percentiles obtained by Kanzler (1990), and therefore we do not expect any improvement when the sample size increases. This is in agreement with the performance of this test, as shown in the table. Regarding the model selection criteria, only the BIC criterion has an acceptable performance. AIC finds non linear structure when it does not exit around one out of four times, and the AICC, although has better performance than AIC, also presents a bad size, especially when the sample size grows. This is in agreement with the fact that BIC is a consistent criterion, and, therefore, the probability of selecting the true model goes to one when the sample size goes to infinity. The AIC and AICC are not consistent and we cannot recommend them for detecting non-linearity as the probability of type I error cannot be controlled and grows with the sample size. Note that the consistency property of the BIC criterion leads to an improvement of its performance with larger sample sizes. For instance, in samples of size 250 the BIC has only a type I error of rejecting linearity for linear processes of 1.8%. A conclusion of this table is that only consistent criteria are expected to be useful for testing non linearity. Thus, we have decided to include only the results of the BIC criterion in the following tables.

Tables 3, 4 and 5 indicate the power of the tests and the performance of the BIC criterion in finding non linear behavior. To simplify the interpretation of these

	T	F^5_{Tsay}	BDS_2	BDS_3	BDS_4	Q_{ML}	D_m	AIC	BIC	AICC
M10a	50	0.043	0.059	0.050	0.053	0.031	0.037	0.246	0.041	0.061
	100	0.048	0.059	0.052	0.054	0.045	0.046	0.263	0.032	0.093
	250	0.050	0.046	0.055	0.048	0.051	0.049	0.269	0.018	0.102
Average		0.047	0.055	0.053	0.052	0.042	0.044	0.260	0.030	0.085
M10b	50	0.058	0.070	0.070	0.066	0.028	0.038	0.251	0.052	0.080
	100	0.047	0.064	0.071	0.081	0.039	0.043	0.256	0.037	0.091
	250	0.047	0.049	0.054	0.055	0.046	0.045	0.260	0.020	0.102
Average		0.051	0.061	0.065	0.067	0.038	0.042	0.256	0.036	0.091

Table 2: Size of the tests and type I error of the model selection criteria

tables we have also displayed the estimated main effects when we consider each table as presenting the output of an ANOVA experiment with two factors: model and method. Thus, the estimation of the main effect for a particular method is obtained as the difference between the average power of all the methods and the average power for this particular one. Let $\overline{y}_{i.}$ be the average power of each method in the nine models considered in the experiment. The main effect of each method is computed as

$$\alpha_i = \overline{y}_{i.} - \overline{y}_{..}$$

where $\overline{y}_{..}$ is the overall mean for all the methods. In the same way, the column $\overline{y}_{.j}$

represents the average power for this model over all the methods and the main effect of each model is estimated by $\beta_j = \overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}$.

Table 3 gives the power of the tests as a function of the model for small sample size, $n = 50$, and the estimated main effects of model and method. The most powerful method is the BIC criterion, with the largest α_i value, $\alpha_{BIC} = .059$, followed by F_{Tsay}^5 with $\alpha_{Tsay} = .029$. The two tests, D_m and BDS_3 have a similar performance, whereas Q_{ML} is clearly behind. All the methods have very small power to detect non linearity in the threshold models, M2 and M3, and in the non linear moving average model, M4.

Tables 4 and 5 present the results for $T = 100$ and $T = 250$. With these larger sample sizes the relative performance of the BDS test, with $k = 3, 4$, improves. This test has the highest average power in tables 4 and 5. For medium sample size, $T = 100$, the average power of BDS_4 is .450, and $\alpha_{BDS4} = .034$, which means that this test has 3.4 points more power than the average of all methods. The F_{Tsay}^5 and the BIC criterion have a good performance, similar to the BDS_3 test. The lowest power corresponds to Q_{ML}, which is a clear loser with relation to all the other methods. For large sample size, $T = 250$, Table 5 shows that BDS_4 is again the best but with small difference with regards to BIC and D_m methods. These methods appear now as second best, with a similar power to the BDS_3 test and very close to the power of BDS_4.

	F^5_{Tsay}	BDS_2	BDS_3	BDS_4	Q_{ML}	D_m	BIC	$\overline{y}_{\cdot j}$	β_j
M1	0.534	0.563	0.581	0.555	0.332	0.478	**0.614**	0.522	0.298
M2	0.069	0.063	0.056	0.055	0.036	0.064	**0.086**	0.061	-0.164
M3	0.055	**0.056**	0.052	0.048	0.040	0.054	0.042	0.049	-0.175
M4	**0.085**	0.037	0.042	0.047	0.035	0.058	0.063	0.052	0.052
M5	**0.200**	0.104	0.133	0.127	0.093	0.150	0.182	0.141	-0.083
M6	0.428	0.326	0.324	0.303	0.191	0.328	**0.462**	0.337	0.113
M7	**0.418**	0.319	0.334	0.319	0.186	0.297	0.386	0.323	0.098
M8	0.121	0.194	0.177	0.158	0.112	0.227	**0.316**	0.187	-0.038
M9	0.372	0.279	0.365	0.403	0.258	0.359	**0.406**	0.349	0.124
$\overline{y}_{i\cdot}$	0.254	0.216	0.229	0.224	0.142	0.224	**0.284**	0.225	
α_i	0.029	-0.009	0.005	-0.001	-0.082	-0.001	**0.059**		

Table 3: Powers for the models in the first experiment when $T = 50$.

We conclude from this first experiment that: (i) the BDS test has the overall better performance, being the winner for medium or large samples, although its power decreases for small samples, (ii) the BIC criterion appears as a strong competitor for the BDS test. It has smaller type I error than the BDS test and better power for small sample size. Also it has only a small difference in power with the BDS test for large sample size; (iii) F^5_{Tsay} and D_m have a overall comparable performance and are slightly behind the BDS and BIC methods. F^5_{Tsay} is better than BDS for small samples, but worse than BIC in this case and D_m is better than BIC for large sample sized, but behind BDS. (iv) Q_{ML} is dominated by the other alternatives.

The second experiment is designed to analyze non linearity in the variance function, such as ARCH, GARCH and stochastic volatility effects. The four models considered are presented in Table 6. M11 corresponds to an ARCH(p) and the parameters a_i have been sampled from an uniform $U(0,1)$ and are re-escaled by an auxiliary variable, s, from an uniform distribution $U(0,1)$ so that $\sum_{i=1}^{p} a_i = s$. M12 and M13 are GARCH(1,1) models with parameters values taken from environmental data (see Tol, 1996), and M14 is a stochastic volatility model from Harvey and Streibel (1998) with $CV(\sigma_t)^2 = .5$. In this experiment two additional tests for heteroskedasticity, are included. The first one was proposed by Harvey and Streibel (1998) and uses the statistic

$$HS = -T^{-1} \sum_{k=1}^{T-1} k r_k,$$

	F^5_{Tsay}	BDS_2	BDS_3	BDS_4	Q_{ML}	D_m	BIC	$\overline{y}_{\cdot j}$	β_j
M1	0.757	0.930	**0.951**	0.943	0.610	0.736	0.824	0.822	0.406
M2	0.082	0.077	0.067	0.068	0.052	0.080	**0.090**	0.074	-0.343
M3	**0.090**	0.065	0.058	0.059	0.047	0.058	0.038	0.059	-0.357
M4	**0.108**	0.040	0.070	0.086	0.062	0.097	0.082	0.078	0.078
M5	**0.606**	0.188	0.280	0.299	0.208	0.292	0.288	0.309	-0.107
M6	**0.860**	0.756	0.757	0.725	0.482	0.669	0.782	0.719	0.303
M7	0.707	0.715	**0.768**	0.767	0.433	0.562	0.672	0.661	0.244
M8	0.189	0.432	0.409	0.377	0.273	0.448	**0.560**	0.384	-0.032
M9	0.532	0.618	0.771	**0.838**	0.501	0.604	0.620	0.641	0.224
$\overline{y}_{i\cdot}$	0.442	0.402	0.445	**0.450**	0.287	0.390	0.435	0.416	
α_i	0.026	-0.014	0.029	**0.034**	-0.130	-0.026	0.019		

Table 4: Powers for the models in the first experiment when $T = 100$.

	F^5_{Tsay}	BDS_2	BDS_3	BDS_4	Q_{ML}	D_m	BIC	$\overline{y}_{\cdot j}$	β_j
M1	0.917	**1.000**	**1.000**	**1.000**	0.910	0.957	0.979	0.966	0.357
M2	**0.151**	0.120	0.104	0.089	0.093	0.137	0.150	0.121	-0.489
M3	**0.209**	0.069	0.065	0.056	0.056	0.064	0.030	0.078	-0.531
M4	**0.196**	0.047	0.137	0.176	0.135	0.185	0.131	0.144	0.144
M5	**0.988**	0.363	0.634	0.669	0.493	0.642	0.575	0.623	0.014
M6	0.993	**0.997**	**0.997**	0.991	0.924	0.981	0.995	0.982	0.373
M7	0.952	0.992	**0.997**	0.996	0.845	0.917	0.960	0.951	0.342
M8	0.274	0.810	0.787	0.738	0.666	0.825	**0.895**	0.713	0.104
M9	0.705	0.960	0.995	**0.999**	0.852	0.908	0.910	0.904	0.295
$\overline{y}_{i\cdot}$	0.617	0.605	0.659	**0.661**	0.567	0.646	0.642	0.609	
α_i	0.008	-0.004	0.049	**0.052**	-0.042	0.037	0.033		

Table 5: Powers for the models in the first experiment when $T = 250$.

M11:	$Y_t = \varepsilon_t \sigma_t$,	$\sigma_t^2 = a_0 + a_1 y_{t-1}^2 + \cdots + a_p y_{t-p}^2$
M12:	$Y_t = \varepsilon_t \sigma_t$,	$\sigma_t^2 = 1.21 + 0.404 y_{t-1}^2 + 0.153 \sigma_{t-1}^2$
M13:	$Y_t = \varepsilon_t \sigma_t$,	$\sigma_t^2 = 1.58 + 0.55 y_{t-1}^2 + 0.105 \sigma_{t-1}^2$
M14:	$y_t = \varepsilon_t \sigma_t$,	$\log(\sigma_t^2) = \mu + \phi \log(\sigma_{t-1}^2) + \eta_t$
	where $\varepsilon_t \sim NID(0,1)$ and $\eta_t \sim NID(0,(1-\phi^2)\sigma_h^2)$.	

Table 6: Models for the second experiment.

where r_k is the autocorrelation coefficients of the squared residuals defined by (4). Under linearity, this statistic has asymptotically the Cramér-von Mises distribution. The second one was proposed by Rodriguez and Ruiz (2005) and uses the statistic

$$RR_m = T \sum_{k=1}^{m-i} \left[\sum_{l=0}^{i} r_{k+l} \right]^2,$$

where $i = m/3 + 1$ and $m = [\sqrt{T}]$. Under linearity it follows a gamma distribution.

The results of this experiment are given in Table 7. For simplicity only the results for $T = 250$ are reported. As before the results are displayed as an ANOVA experiments with two factors: model and tests. The two tests, D_m and BDS_4 have a similar performance and are the best procedures in average power. The BIC criterion has an intermediate performance, similar to Q_{ML} and RR_m, but much better than HS and F_{Tsay}^5. When the persistence in the autocorrelations increases, the average power in the BIC and BDS decreases (see M11 with $p = 5$ and M14 with $\phi = 0.985$) and HS and RR_m have a good performance in SV with ϕ close to 1 where the autocorrelation

		F^5_{Tsay}	BDS_2	BDS_3	BDS_4	Q_{ML}	D_m	RR_m	HS	BIC	$\overline{y}_{\cdot j}$	β_j
M11	p=1	0.413	0.812	0.796	0.771	0.716	0.764	0.589	0.261	0.796	0.658	0.056
	p=3	0.411	0.552	0.649	0.695	0.634	0.668	0.638	0.347	0.650	0.583	-0.019
	p=5	0.380	0.445	0.525	0.574	0.585	0.601	0.618	0.391	0.535	0.517	-0.084
M12		0.393	0.959	0.962	0.951	0.846	0.920	0.666	0.253	0.940	0.766	0.164
M13		0.552	0.994	0.995	0.989	0.959	0.984	0.825	0.352	0.984	0.848	0.247
M14	0.840	0.247	0.365	0.482	0.544	0.395	0.457	0.449	0.247	0.341	0.392	-0.209
	0.985	0.233	0.274	0.374	0.433	0.571	0.527	0.658	0.646	0.305	0.447	-0.155
$\overline{y}_{i\cdot}$		0.376	0.629	0.683	0.708	0.672	0.703	0.635	0.357	0.650	0.601	
α_i		-0.226	0.027	0.082	0.107	0.071	0.101	0.033	-0.245	0.049		

Table 7: Powers for the models in the second experiment when $T = 250$.

is very persistent (M14 with $\phi = 0.985$).

The third experiment follows the design by Barnett et al. (1997). Five models were included in this study. The first is the logistic equation or the deterministic chaotic Feigenbaum sequence:

$$y_t = ay_{t-1}(1 - y_{t-1})$$

with $a = 3.57$ and initial condition $y_0 = .7$. The logistic equation may produce sample paths looking as a nonstationary process or white noise depending on a and y_0. The second model is a GARCH(1,1). The third is the nonlinear moving average,

$$y_t = \epsilon_t + .8\epsilon_{t-1}\epsilon_{t-2},$$

the fourth is an ARCH(1) and, finally, the fifth is an ARMA (2,1) model. Table 8 shows the results for a sample size $n = 380$. Following the design in Barnett et al. (1997), we report if the hypotheses of linearity was accepted (A), rejected (R) or the procedure was ambiguous (?) about it. The table presents the results of the four previous tests, the BIC criterion and four of the tests included in the study by Barnett et al. (1997) with the same samples generated by these five models. Regarding the BDS we have taken the implementation used in Barnett et al. (1997). From Table 8 it can be seen that the only procedure which always finds the correct answer is the Kaplan test. The second best are the D_m and Q_{ML} tests and the BIC criterion, which does not detect the non linearity in the GARCH model. Table 9 shows the results for $n = 2000$, and now four procedures, Kaplan, D_m, Q_{ML}, and the BIC criterion, are able to find the right answer.

The conclusion of these three experiment is that, among the tests compared, the BDS test performs the best. It has overall the highest power for detecting non linear behavior in both the mean and the variance function of the process and also seems to be able to detect chaotic behavior. The second best among the tests compared is the D_m. It is similar to the F^5_{Tsay} for the mean, but clearly superior for detecting non linearity in the variance function. Also, it detects the chaotic behavior in the

Process	F^5_{Tsay}	Q_{ML}	D_m	BIC	Hinich	BDS	White	Kaplan	True
I (Feig)	R	R	R	R	A	R	R	R	R
II (GARCH)	A	A	A	A	A	?	A	R	R
III (NLMA)	R	R	R	R	A	?	A	R	R
IV (ARCH)	?	R	R	R	A	?	A	R	R
V (ARMA)	A	A	A	A	A	A	A	A	A

Table 8: Models from Barnett et al, A=accept linearity, R=Reject, ?=Ambiguous, T=380

Process	F^5_{Tsay}	Q_{ML}	D_m	BIC	Hinich	BDS	White	Kaplan	True
I (Feig)	R	R	R	R	A	R	R	R	R
II (GARCH)	A	R	R	R	A	?	A	R	R
III (NLMA)	R	R	R	R	A	?	A	R	R
IV (ARCH)	?	R	R	R	A	?	A	R	R
V (ARMA)	A	A	A	A	A	A	A	A	A

Table 9: Models from Barnett at al, A=accept linearity, R=Reject, ?=Ambiguous, T=2000

third experiment. The BIC criterion is a strong competitor of the best linear test in all cases:it is better than BDS and D_m for small samples sizes and it has almost the same power for large sample.

5 An example

We will explore the non linearity in the series of quarterly US real GNP, Y_t, from the first quarter of 1947 to the second quarter of 2003. The data are seasonally adjusted and are shown in Figure 1. Figure 2 shows the rate of growth of this series given by the transformation $y_t = \nabla \log Y_t$. This series has been extensively analyzed in the econometrics and statistics literature, see for instance Tiao and Tsay (1994), and also in the economic literature, see for instance McConnell and Pérez-Quirós (2000).

The best ARMA model fitted for this series, as selected by BIC, is model M_1. The second row of Table 10, gives the estimated parameter values, the BIC and the Ljung-Box statistics, $Q_{LB}(10)$, for this model M_1. The residuals of this model show some extreme values which can be modelled as outliers and this leads to model M_2. The third row of Table 10 describes this model which includes two additive outliers (AO) a transitory change (TC) and one level shift (LS), as detected by program TSW, Windows version of TRAMO-SEATS (© Gómez and Maravall, 1996). The table shows that, as expected, model M_2 with outlier correction has a smaller residual variance and a smaller value of BIC. We have applied to the residuals of these two

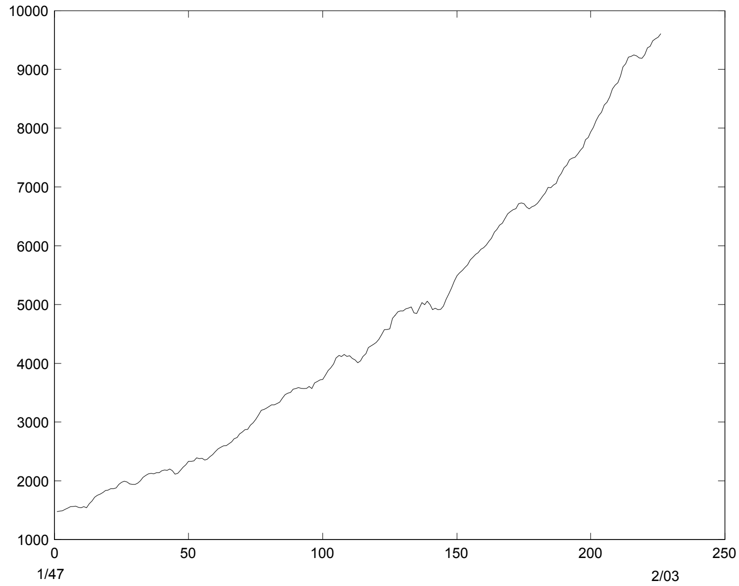


Figure 1: Real US GNP seasonally adjusted from 1947 to 2003.

models the previous non linearity tests and the outcomes are presented in the second and third rows of Table 11. As most of the tests detect non linear behavior in the residuals of these two models, we conclude that the series seems to be non linear. We also note that M_2 is found non linear more often than M_1, which implies that cleaning this series from outliers makes easier the identification of the non linear behavior. The residuals and the autocorrelation function of the squared residuals from model M_2 are given in Figure 3. Note that, among the first 20 autocorrelations, those of lags 2, 4, 6 and 9 seem to be different from zero.

As the sample is large, one possible explanation of the detected non linear behavior is the presence of a structural break in the period. In order to explore this

Model	Period	Size	AR	MA	BIC	$\widehat{\sigma}^2 \times 10^{-4}$	AO	TC	LS	$Q_{LB}(10)$
M_1	1/47-2/03	226	0.342 (0.063)	–	-9.28	0.900	–	–	–	11.39
M_2	1/47-2/03	226	0.423 (0.060)	–	-9.46	0.693	4/49	1/58	2/78	12.57
							4/70			
M_3	1/47-1/75	113	0.468 (0.083)	–	-9.09	0.989	4/49	–	–	8.26
							4/70			
M_4	2/75-2/03	113	0.477 (0.083)	–	-10.01	0.380	2/81	–	2/78	9.54
									2/80	
M_5	1/47-1/83	145	0.429 (0.075)	–	-9.07	1.001	4/49	–	2/78	7.59
							4/70			
M_6	2/83-2/03	81	–	-0.306 (0.106)	-10.44	0.258	–	–	–	8.4
				-0.352 (0.107)						
M_7	4/83-2/03	79		-0.283 (0.11)	-10.49	0.247	–	–	–	6.11
				-0.318 (0.11)						

Table 10: Models for US Real Growth series in different periods.

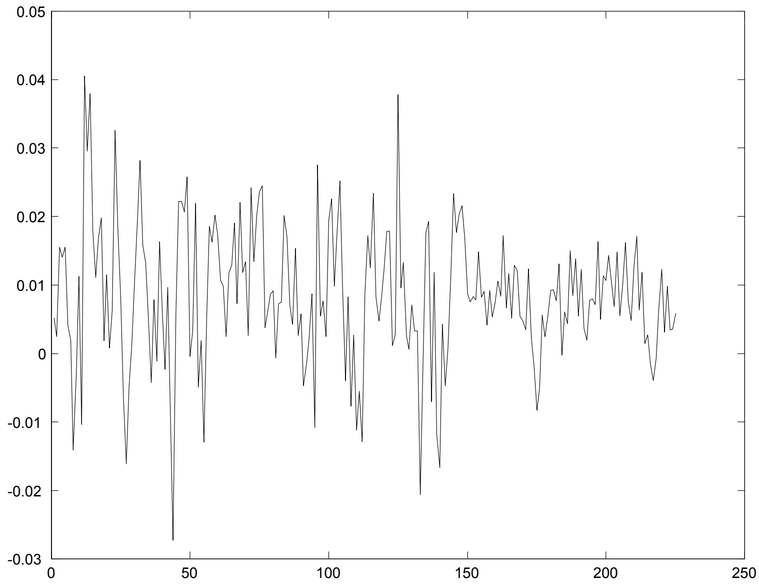


Figure 2: Rate of growth of US GNP. 1947 to 2003

possibility we split the series into two halves and analyze the non-linear behavior in each subsample. The first half is taken from 1/47-1/75, with 113 observations, and the second from 2/75-2/03, also with 113 observations. The models fitted to the two subsamples are M_3 and M_4, and are given in rows 4th and 5th of Table 10. We see that the estimated AR parameter has a similar value in both models, whereas the residual variance is much smaller in M_4. In fact, the standard F test of comparison of the variances for the two models is highly significant. The results of the linearity tests are given in rows 4th and 5th of Table 11. We found that all the tests indicate that the series is linear in the first half period, and non linear in the second half. This non linear behavior happens with a strong reduction of variability, as the resid-

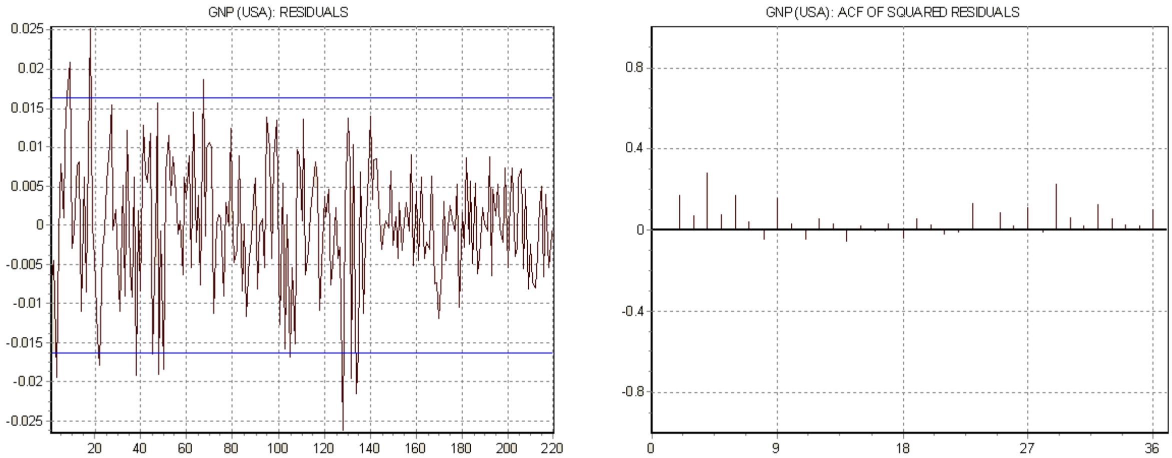


Figure 3: Residuals and the acf of these squared residuals from model M_2.

ual variance of M_4 is one third of the one of M_3. In order to understand better this change in variability, Figure 4 shows a robust measure of scatter, the MAD (median of absolute deviations), computed in non overlapping groups of 24 observations (6 years) over the whole period. This figure indicates that the variance in the last three groups, which correspond to the last 18 years in the sample, is much smaller than in the rest of the groups.

In order to identify the time of this variance change in the series we apply the Cusum procedure for retrospective detection of variance changes developed by Inclán and Tiao (1994). The plot of the statistic proposed by these authors is given in Figure 5. This statistic shows that the largest change is around observation 145. Note that Figure 5 gives, in a more accurate way, the same information that was found in Figure 4: a decrease in variability after observation 50, followed by a larger decrease around

Model	Period	BIC	D_m	Q_{ML}	$F_{Tsay}^{2,3,4,5}$	$BDS_{2,3,4,5}$	RR	HS
M_1	1/47-2/03(NO)	L	NL	NL	L,L,L,L	L,NL,NL,NL	NL	NL
M_2	1/47-2/03	NL	NL	NL	L,L,L,L	L,L,L,NL	NL	NL
M_3	1/47-1/75	L	L	L	L,L,L,L	L,L,L,L	L	L
M_4	2/75-2/03	NL	NL	NL	NL,NL,NL,NL	NL,NL,NL,NL	NL	NL
M_5	1/47-1/83	L	L	L	L,L,L,L	L,L,L,L	L	L
M_6	2/83-2/03	NL	L	L	NL,NL,NL,NL	L,L,L,L	L	L
M_7	4/83-2/03	L	L	L	NL,NL,NL,NL	L,L,L,L	L	L

Table 11: Non linearity tests applied to the residuals of models for US Real Growth series in different periods.

observation 145. These findings are also in agreement with the residuals plot in Figure 3. From now on we will concentrate in this large variance change at $t = 145$.

As the large variance change occurs in the time period used to fit model M_4, which showed non linear behavior, we wonder if this non linearity can be due to the variance change. In fact, some authors, see for instance Shumway and Stoffer (2002), have found GARCH effects in this series. We show in the appendix that a variance change in a residual series is expected to produce correlations among the squared residuals. Thus, the large variance change found in this series can be responsible for the non linear behavior observed when checking the autocorrelations of the squares. As it is well known that squared autocorrelations could also be due to the conditional

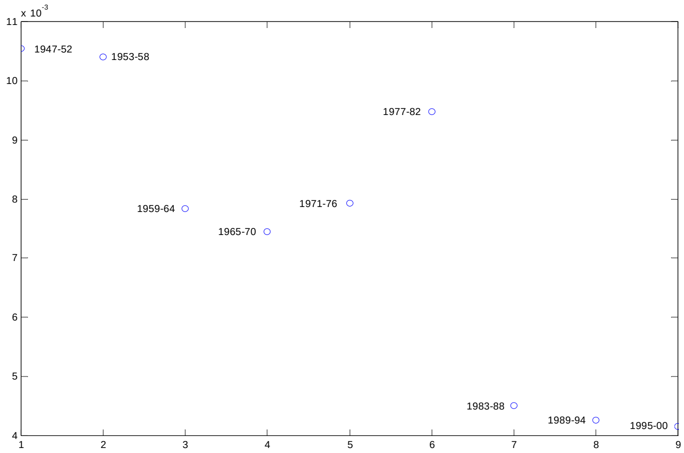


Figure 4: Median absolute deviation (MAD) and time in groups of 24 observation for the real growth GNP.

heteroskedasticity of a GARCH model, in order to differentiate between these two explanations we have fitted a GARCH model and a variance change model to the residuals e_t, of the GNP series in the whole period 1/1947-2/2003, after cleaning these residuals from outliers. The estimated GARCH model is

$$\begin{aligned} e_t &= \varepsilon_t \sigma_t \\ \sigma_t^2 &= 4.8 \times 10^{-7} + 0.9292\sigma_{t-1}^2 + 0.0601 e_{t-1}^2 \end{aligned}$$

and Figure 6 shows the squared residuals and the estimated volatility σ_t^2. A global measure of the fit for this model is given by

$$\sum \frac{(e_i^2 - \sigma_t^2)}{T} = 1.0961 \times 10^{-8}.$$

We have compared this measure to the one obtained with the variance change model

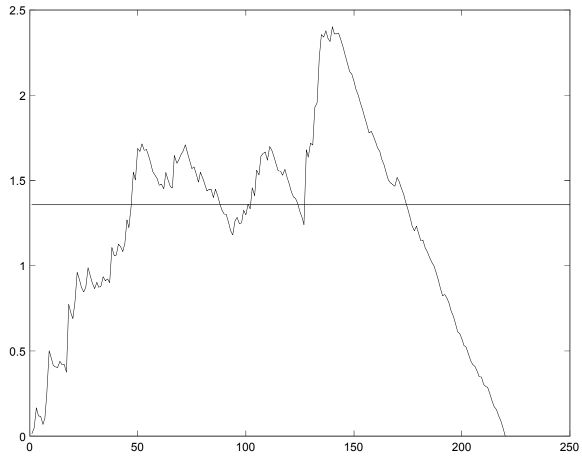


Figure 5: Cusum chart for identifying variance changes.

given by

$$\begin{aligned} e_t &= \varepsilon_t \sigma_t \\ \sigma_t^2 &= 1.0 + (.26 - 1.0) S_t^{(145)} \end{aligned}$$

where $S_t^{(145)}$ is a step function. Note that this model implies that for $t < 145$ the residual variance is 1.001 and for $t \geq 145$ the residual variance is .258. Figure 7 shows the plot of the residuals from this model which has a measure of fit of

$$\sum \frac{(e_i^2 - \sigma_t^2)}{T} = 1.0334 \times 10^{-8}.$$

We conclude that, although both models seem to be compatible with the data, the variance change model gives a better fit with smaller number of parameters. Thus, it will be the one selected by any model selection criterion.

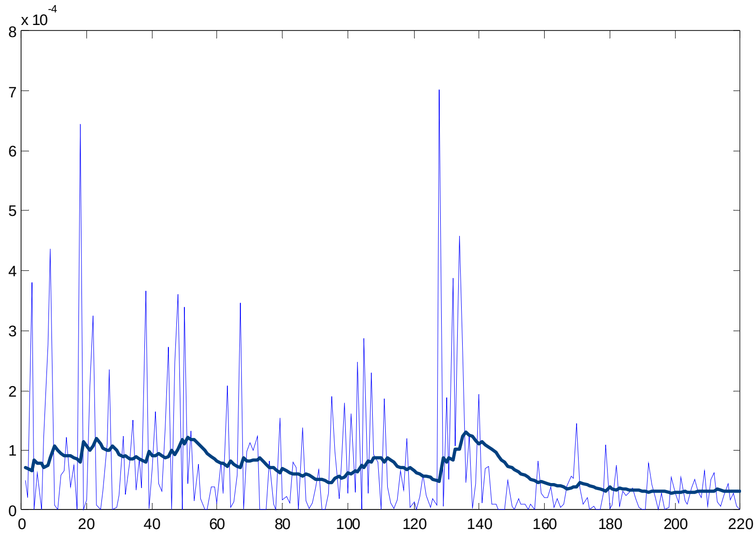


Figure 6: Squared residuals and conditional variance for the GARCH model estimated to the residuals of the GNP series. Period 1/1947-2/2003.

The previous analysis suggests that instead of splitting the series into two halves, it may be more informative splitting it before and after the large variance change. Thus, we now split the total available time period in the subsamples 1/47 to 1/83 and from 2/83 to 2/03, and estimate models M_5 and M_6, given in rows 5th and 6th of Table 10, in these two periods. The best model fitted in the first subsample (1/47 to 1/83) according to BIC is an AR(1), see model M_5, whereas in the second period is a MA(2), see model M_6. The residual variances estimated in both periods are very different and similar to the ones estimated by the variance change model. The results of the non linearity tests applied to the residuals from models M_5 and M_6 are given

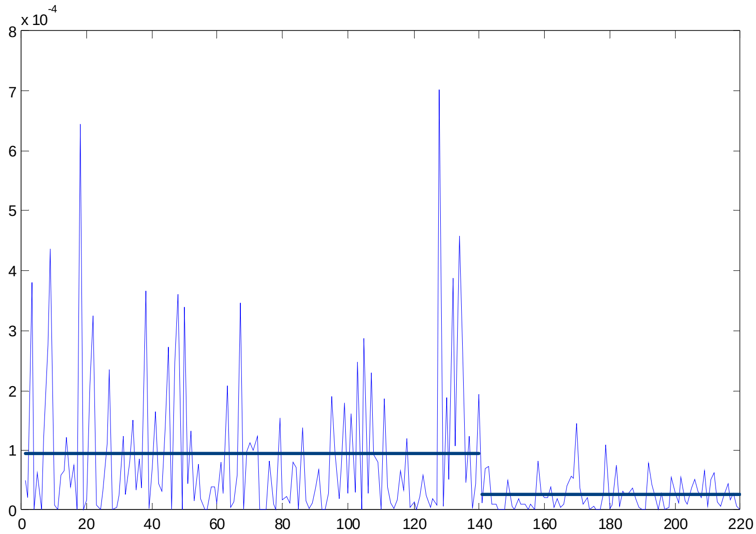


Figure 7: Squared residuals and variance change in the residuals of the GNP series. Period 1/1947-2/2003.

in rows 5th and 6th of of Table 11. Now the first series is clearly linear, whereas the last one is unclear: The BIC and Tsay tests indicate non linear behavior, whereas the other tests, based on the correlation of the squared residuals, do not reject the linearity hypotheses. We conclude that the strong non linear behavior found in model M_4 was probably due to the large variance change in the time period used to fit this model.

Let us study with more detail the posibility of non linear behavior in the second subsample, from 2/83 to 2/03, by comparing the performance of nonlinearity tests for the residuals of model M_6 . Figure 8 presents the residual plot and the autocorrelation function of the squared residuals.There is a relatively large coefficient at lag 1: $r_1(\varepsilon_t^2) = 0.3191$ with a standard deviation of 0.1125. This explains the non linear behavior of M_6 found in Table 11 by the BIC criterion, as an AR(1) model will be appropriate for the series of squared residuals. The failure of the Q_m and the D_m tests in rejecting the linearity hypotheses is due to the large value of m chosen. We are always using $m = \sqrt{T}$, which means $m = 9$ in this case, and as we only have a significant coefficient these tests will have power for a small value of m, as for instance $m = 5$, but not for larger m.

We have carried out a sensitivity analysis for checking the influence of the splitting time in the presence of non linearity in the last part of the series. As the starting point of the period of smaller variance is not clear, as indicated in Figure 5, we have also analyzed the second subsample but starting at 4/83 instead of 2/83. This leads to model M_7, presented in the last rows of Tables 10 and 11. When comparing M_6 and M_7, the estimated parameters change slightly and the results of the tests in Table 11 are the same, expect for the BIC test. For M_6 the BIC indicates non linear behavior whereas for M_7 it accepts linearity. Thus, only the Tsay test keeps showing indication of non linear behavior for both models M_6 and M_7, and the reason for this is a strongly significant coefficient associated to ε_{t-1}^2. The plots of ε_t with respect to both ε_{t-1} and ε_{t-1}^2 are shown in Figure 9. Both plots shows clear signs of threshold autoregressive (TAR) non linear behavior. Figure 9a shows that the dependency between ε_t and ε_{t-1} seems to be different when $\varepsilon_{t-1} < 0$ and when $\varepsilon_{t-1} > 0$. The plot of ε_t with respect ε_{t-1}^2 shows a negative relationship between both variables, which explains why the Tsay detects non linearity. As most of the portmanteau non linearity test are not powerful for TAR, it is not surprising that they fail to find this type of non linear behavior. It is interesting to note that Tiao and Tsay (1994) found evidence of threshold behavior in this series and fitted a four regimes TAR model to data in the period 1/47 to 1/91. We conclude that the series has very likely TAR behavior and only the Tsay test, among the tests considered, has been able to show this feature. The estimated TAR model on the original data is:

$$M_8: \qquad y_t = \begin{cases} \underset{(.001)}{.0031} + \underset{(.1711)}{1.0345} y_{t-1} + \varepsilon_t^1 & y_{t-1} \leq .0084 \text{ and } \varepsilon_t^1 \sim N(0, 1.87 \cdot 10^{-5}) \\ \underset{(.003)}{.0045} + \underset{(.2525)}{.2340} y_{t-1} + \varepsilon_t^1 & y_{t-1} > .0084 \text{ and } \varepsilon_t^1 \sim N(0, 1.77 \cdot 10^{-5}) \end{cases}$$

We conclude that the series of US real GNP growth has suffered a structural break in 1983. Before this period the series was linear and follows an AR(1). After 1983, the variance is reduced to 1/4 and the series shows non linear TAR behavior.

Finally, we have analyzed the out of sample performance of models M_1, M_2, M_7 and M_8. The last ten observations in the time series have been dropped, the model estimated without them, and ten one-step-ahead out-of-sample forecasts have been computed by rolling forecasts, that is reestimating the model when a new observation becomes available for the next one-step-ahead forecast. The mean squared prediction error of these four models for one step ahead out of sample forecasts are given in

M_1	M_2	M_7	M_8
5.42	1.67	2.63	4.66

Table 12: Mean squared prediction error of four models for one-step-ahead out of sample forecasts. The values given are divided by 10^5

Table 12

It can be seen that the TAR model, M_8, does not improve the forecasts obtained by models M_2 and M_7 in this exercise. These two models are both linear, and this result confirms that, as often found by other authors, non linear models may not provide clear gains over liner ones in out of sample forecast. Also, allowing for outliers and level shifts in a linear model, as in model M_2, can lead to a significative forecast improvement with respect to a linear model as M_1 which does not take them into account.

6 Conclusion

The main conclusion from this paper is that by checking with the BIC criterion if the order selected when fitting an AR model to the squared residuals of a linear fit is zero we may have an effective way for detecting non linear behavior. The BIC criterion has an overall good performance: its power for detecting non linearity is either the largest or close to the first best of the tests compared. For large sample size the type I error of the BIC is the smallest among the procedures compared.

Also this procedure is robust to the parameter $p_{\max}$, maximum order of the AR fitted to the squared residuals. The other information criteria considered cannot be recommended, because of their large type I error. Thus, efficient criteria do not seem to be useful with this objective, whereas the BIC property of consistency guarantees a good performance in large samples. The worst behavior of the BIC criterion is found for detecting some forms of heteroskedasticity with respect to tests designed to take into account the expected structure of the squared autocorrelations. Also, it has no power for threshold behavior. Therefore, we conclude that although the BIC criterion is useful as a kind of Portmanteau non linearity test, it is better to supplement it with specific tests for the type of non linear behavior that is expected to appear in the data.

The BDS test has also an overall good performance, confirming the results obtained in previous studies. The F_{Tsay}^{5} and D_m tests are simpler alternatives which can work as well, or better, than BDS in small samples and are competitive in large samples. In particular, as shown in the example, the F_{Tsay}^{5} is able to show threshold behavior in situations which are non detected by the rest of the tests included in our study.

A conclusion we draw from the example is that we should be careful when interpreting the results of a test that finds significative autocorrelation among the squared residuals. This could be due to a non linear model, outliers, variance changes or conditional heteroskedastic models, and it is important to differentiate among these effects. Finally, taking these changes into account can have a large improvement in the forecasting performance of the model.

Acknowledgement We are very grateful to Antonio García Ferrer, Esther Ruiz and Pilar Poncela for many useful comments which have improved the paper very much. Antonio provides the data for the example and gave many suggestions which motivated us to do a deeper analysis of this data set. We are also grateful to Pedro Galeano for providing us with his software for estimating TAR models. This research has been supported by MEC grant SEJ2004-03303 and the Fundación BBVA.

Appendix

We compute the large sample autocorrelations of the squared values in a white noise series, a_t, with a variance change. Suppose that the variance change happens at time $t = h = \alpha T$, where to simplify we assume that αT is an integer, T is the sample size and $\alpha \in (0, 1)$ and, without loss of generality, let us assume that the variance changes from 1 to c^2. Thus, from $t = 1, ..., h$ we observe a_t with variance 1 and from $t = h + 1, ..., T$ we observe ca_t. Then, the variance of the series will be

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{h} a_t^2 + c^2 \sum_{t=h+1}^{T} a_t^2}{T}$$

and, assuming T large, we approximate this variance for its expected value,

$$\sigma^2 = \alpha + c^2(1 - \alpha)$$

The autocorrelation of the squares are

$$r_k = \frac{\sum_{t=k+1}^{T}\left(a_t^2-\hat{\sigma}^2\right)\left(a_{t-k}^2-\hat{\sigma}^2\right)}{\sum_{t=1}^{T}\left(a_t^2-\hat{\sigma}^2\right)^2}$$

and the numerator can be written as

$$\sum_{t=k+1}^{T}\left(a_t^2-\hat{\sigma}^2\right)\left(a_{t-k}^2-\hat{\sigma}^2\right) = \sum_{t=k+1}^{T} a_t^2 a_{t-k}^2 - \hat{\sigma}^2\sum_{t=k+1}^{T} a_t^2 - \hat{\sigma}^2\sum_{t=k+1}^{T} a_{t-k}^2 + (T-k)\hat{\sigma}^4$$

and if we now approximate each term by its expected value

$$E(\sum_{t=k+1}^{T} a_t^2 a_{t-k}^2) = (h-k) + kc^2 + (T-k-h)c^4$$

and using that, approximately

$$E(\hat{\sigma}^2\sum_{t=k+1}^{T} a_t^2) \approx E(\hat{\sigma}^2\sum_{t=k+1}^{T} a_{t-k}^2) \approx (T-k)\sigma^4$$

we have that

$$E(\frac{1}{T}\sum_{t=k+1}^{T}\left(a_t^2-\hat{\sigma}^2\right)\left(a_{t-k}^2-\hat{\sigma}^2\right)) \approx \alpha + c^4(1-\alpha) - \sigma^4 + (k/T)(\sigma^4 + c^2 - c^4 - 1).$$

The denominator can be approximated by

$$E(\frac{1}{T}\sum_{t=1}^{T}\left(a_t^2-\hat{\sigma}^2\right)^2) = 3(\alpha + c^4(1-\alpha)) - \sigma^4$$

and, therefore

$$r_k = \frac{\alpha + c^4(1-\alpha) - \sigma^4 + (k/T)(\sigma^4 + c^2 - c^4 - 1)}{3(\alpha + c^4(1-\alpha)) - \sigma^4}$$

This expression shows that the autocorrelacion coefficients of the squared values of the time series will be different from zero. For instance, for large c^2, this function can be approximated by

$$r_k = \frac{1}{3} - \frac{(2-\alpha)}{3(1-\alpha)}\frac{k}{T}$$

which is greater than zero, and could be close to 1/3. Note that the autocorrelacion coefficients will decrease slowly, and their structure would be similar to the one produced by GARCH effects.

References

Akaike, H. (1969). Fitting autoregressive models for prediction, *Ann. Inst. Stat. Math.* **21**, 243-247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory* (Akademiai Kiadó, Budapest), 267-281.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control AC* **19**, 203–17.

Ashley R. A. & Patterson D. M. (1999). Nonlinear model specification/diagnostics: Insights from a battery of nonlinearity tests, Working Paper E99-05, *Department of Economics,* Virginia Tech.

Barnett, W. A., Gallant, A. R., Hinich, M. J., Jungeilges, J. A., Kaplan D. T. & Jensen M. J. (1997). A single-blind controlled competition among test for nonlinearity and chaos, *Journal of Econometrics*, **82**, 157-192.

Brock, W., Dechert, D., Scheinkman, J. & LeBaron, B. (1996). A test for independence based on the correlation dimension, *Econometric Reviews,* **15**, 197-235.

Brock, W., Hsieh, D., &, LeBaron, B. (1991). *Nonlinear dynamics, chaos, and instability: Statistical theory and economic evidence.* MIT Press: Cambridge, MA.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variances, *Econometrica*, 50, 987–1007.

Fan, J. & Yao, Q. (2003). *Nonlinear time series: Nonparametric and parametric methods,* Springer.

Galeano P. & Peña, D. (2004). Model selection criteria and quadratic discrimination in time series, *Working Paper n° 04-14, Universidad Carlos III de Madrid,* Madrid.

Gómez V., & Maravall A. (1996). Programs SEATS and TRAMO: Instructions for the user, *Working Paper n° 9628*, Bank of Spain.

Granger, C. W. J. & Andersen, A. P. (1978). *An introduction to bilinear time series models*, Götttingen,Vandenhoeck & Rpurecht.

Granger, C. W. J., & Teräsvirta, T. (1992). *Modeling nonlinear economic relationships,* Oxford University Press.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression, *J. R. Stat. Soc. B,* **41**, 190-195.

Hannan, E. J. & Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order, *Biometrika,* **69**, 81-94.

Harvill, J. L. (1999). Testing time series linearity via goodness-of-fit methods, *Journal of Statistical Planning and Inference*, Vol. 75, **2**, pp. 331 - 341.

Harvey, A. & Streibel, M. (1998). Testing for slowly changing level with special reference to stochastic volatility, *Journal of Econometrics*, **87**, 167-189.

Hjellvik, V. & Tjϕstheim, D. (1995). Nonparametric tests of linearity for time series, *Biometrika*, 82, 351-368.

Hinich, M. J. (1982). Testing for gaussianity and linearity of a stationary time series, *J. Time Series Anal.* **3**, 169-176.

Hurvich, C. M. & Tsai, C. (1989). Regression and time series model selection in small samples, *Biometrika,* **76**, 297-307.

Inclán, C. & Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance, *Journal of the American Statistical Association,* **89**, 913–923.

Kaplan, D. T. (1994). Exceptional events as evidence for determinism, *Phisyca D*, 73, 38-48.

Kanzler, L. (1999). Very fast and correctly sized estimation of the BDS statistic, *Department of Economics of Oxford University.* http:\users.ox.ac.uk/∼econlrk

Keenan, D. M. (1985). A Tukey Nonadditivity-type test for time series nonlinearity, *Biometrika* 72, 39–44.

Koreisha, S. & Pukilla, T. (1995). A comparison between different order-determination criteria for identification of ARIMA models, *Journal of Business & Economics Statistics*, 13, 127-131.

Lee, T. H., White, H., & Granger, C.W.J. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests, *Journal of Econometrics*, 56, 269-290.

Luukkonen, R., Saikkonen, P. & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models, *Biometrika* 75, 491–499.

Maravall, A. (1983). An application of nonlinear time series forecasting, *Journal of Business & Economics Statistics*, 1, 66–74.

McLeod, A.I. & Li, W.K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations, *Journal of Time Series Analysis*, 4, 269-273.

McConnell. M. M. & Pérez Quirós, G. (2000). Output fluctuations in the United States: What has changed since the early 80s?, *American Economic Review,* 90, 5, 1464-1476.

Monti, A. C. (1994). A proposal for residual autocorrelation test in linear models, *Biometrika,* 81, 776-80.

Peña, D. & Rodríguez, J. (2002). A Powerful Portmanteau Test of Lack of Fit for Time Series, *Journal of the American Statistical Association,* **97**, 601–610.

Peña, D. & Rodríguez, J. (2005). The Log of the determinant of the autocorrelation matrix for testing goodness of fit in time series, *Journal of Statistical Planning and Inference*, (in press).

Peña, D., Tiao, G. C., & Tsay, R. S. (2001). *A Course in Time Series Analysis*, John Wiley & Sons, Inc.

Priestley, M. B. (1989). *Spectral Analysis and Time Series, Vol.1*, Academic Press Inc., San Diego.

Pukkila, T., Koreisha, S.& Kallinen, A. (1990). The identification of ARMA models, *Biometrika*, **77**, 537-548.

Rodríguez, J. & Ruiz E. (2005). A powerful test for conditional heteroscedasticity, *Statistica Sinica,* (in press).

Shumway, R. H. & Stoffer, D. S. (2000). *Time series analysis and its applications,* Springer-Verlag, N. Y.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.* **6**, 461-464.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Ann. Statist.* **8**, 147-164.

Subba Rao, T., & Gabr, M. M. (1980). A Test for linearity of stationary time series, *J. Time Series Anal.* **1**, 145-158

Terdik, G. (1999). *Bilinear stochastic models and related problems of nonlinear time series analysis: A frequency domain approach*, Lecture Notes in Statistics, No. 142, Springer-verlag: New York.

Tiao, G. C. & Tsay, R. S. (1994). Some advances in nonlinear and adaptive modeling in time series (with discussion), *Journal of Forecasting 13*, 109-140.

Tol, R. S. J. (1996). Autoregressive conditional heteroscedasticity in daily temperatures measurements, *Environmetrics,* **7**, 67–75.

Tong, H. (1990). *Nonlinear time series analysis: A dynamical systems approach,* Oxford University Press.

Tsay, R. S. (1986). Nonlinearity tests for time series, *Biometrika*, **73**, 461–466.

Tsay, R. S . (1989). Testing and modelling threshold autoregressive processes, *Journal of the American Statistical Association,* **84**, 2 31-240.

Tsay, R. S. (1991). Nonlinear time series analysis: Diagnostics and modelling, *Statistica Sinica*, **1**, 431-451.

Tsay, R. S. (2001). Nonlinear time series analysis, in *A Course in Time Series Analysis*, Peña, D., Tiao, G. C., & Tsay, R. S. (eds). John Wiley & Sons, Inc.

Biographies

DANIEL PEÑA is Professor of Statistics at Universidad Carlos III de Madrid, Spain. He has published 12 books and more than 150 papers on time series and forecasting, diagnostic and robust methods, Bayesian statistics, multivariate analysis, econometrics and quality methods. He is an ISI member and IMS fellow.

JULIO RODRIGUEZ is Assistant Professor of Statistics at Universidad Politecnica de Madrid, Spain. His areas of interests are time series, multivariate analysis and graphical methods. His reserach work has appeared in the *Journal of American Statistical Association, The Journal of Multivariate Analysis and Statistica Sinica,* among others.

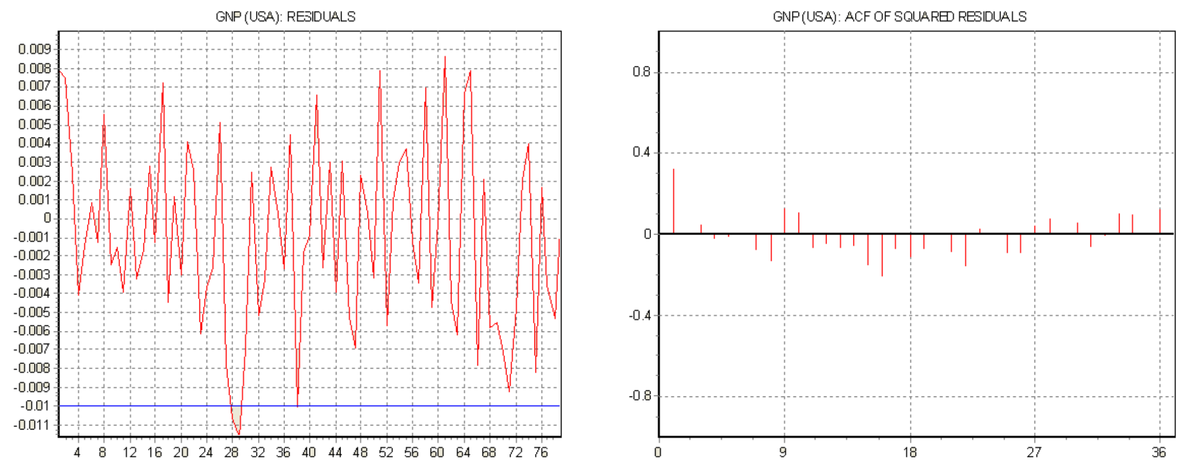
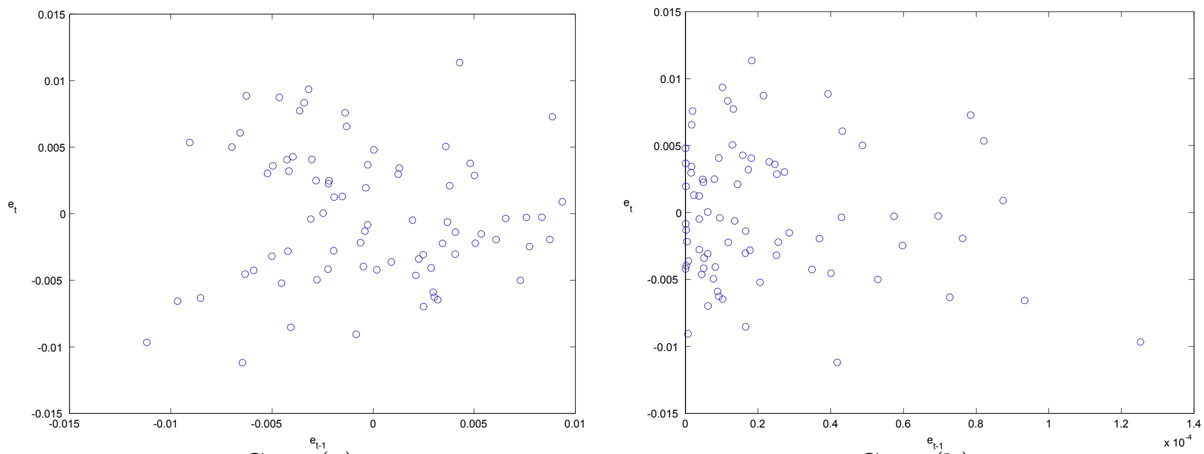


Figure 8: Residuals of their acf of squared residuals for model M_6.



Case (a) Case (b)

Figure 9: Residuals of US GNP with respect to lag residuals (left, case (a)) and to squared lag residuals (right, case (b)). Model M_7.