Bootstrap Power of Time Series Goodness-of-Fit Tests

Sohail Chand
College of Statistical and Actuarial Sciences
University of the Punjab, Q.A. Campus
Lahore, Pakistan
sohail.stat@gmail.com

Shahid Kamal
College of Statistical and Actuarial Sciences
University of the Punjab, Q.A. Campus
Lahore, Pakistan
kamal_shahid@hotmail.com

Abstract

In this article, we looked at power of various versions of Box and Pierce statistic and Cramer von Mises test. An extensive simulation study has been conducted to compare the power of these tests. Algorithms have been provided for the power calculations and comparison has also been made between the semi parametric bootstrap methods used for time series. Results show that Box-Pierce statistic and its various versions have good power against linear time series models but poor power against non linear models while situation reverses for Cramer von Mises test.

MSC: 62M10, 91B84

Keywords: Portmanteau tests; Bootstrapping; Power.

1. Introduction

Time series model building is a science and art as well. It is generally considered a three stage iterative procedure consisting of identification, estimation and diagnostic checking (Box and Jenkins, 2008). Diagnostic checking is an important stage and residuals obtained by fitting the identified model play an important role in model criticism. Box and Pierce (1970) test and its several other versions are perhaps the most commonly used types of portmanteau test (Mainassara et al., 2009). These tests are capable to perform an overall test for an entire set of, say, the first m autocorrelations assuming that the null model, i.e. the model defined under null hypothesis, is correct. Moreover, the choice of m is very important in the appropriateness of asymptotic distribution and power of these tests.

In this paper, we numerically study the power of some of the popular time series goodness of fit tests. Escanciano (2006) has studied power of various goodness of fit tests under the fixed design wild bootstrap. Horowitz et al. (2006) has compared performance of Box and Pierce (1970) test with some other tests under double block bootstrapping.

The novelty of our study is that we study the size of the tests under various semi-parametric bootstrap designs described in Section 3.1. We compare the power of these tests with the Cramer von Mises (CvM) (Escanciano, 2007) statistic against various linear and non-linear alternative models. To the best of our knowledge, these tests are not studied under these setting in the literature.

Our results show that Box-Pierce type tests do well against the linear alternatives but fail to perform against the non-linear alternatives, while the situation reverses for the CvM statistic due to Escanciano (2007), i.e, the CvM statistic does well against various non linear alternatives but much less well against various linear alternatives. Moreover, dynamic bootstrap methods show better performance than the fixed design bootstrap in our example. We have not found any clear advantage of using wild residuals bootstrapping in this scenario.

The remainder of the paper is organized as follows: In the next section a review of the literature on available diagnostic tests is given. Section 3 describes the different bootstrap methods in time series context. Section 4 gives the estimation procedure and algorithms for Monte Carlo simulations for computing power of the tests. Finally, Section 5 presents the results of simulations and discussion of the results.

In practice, there are many possible linear and non-linear models for a problem under study e.g. autoregressive, moving average, mixed ARMA models, threshold autoregressive etc. Box and Jenkins (2008) have described time series model building as a three-stage iterative procedure that consists of identification, estimation and validation.

Identification of the model is partly science and partly arts. There are no exact ways of identifying the underlying model though there are some tools, for example, the autocorrelation and partial autocorrelation plots to identify the general class of underlying model, see Box and Jenkins (2008, p.196). Importantly, it should be noted that at the identification stage, especially dealing with complex situations, we identify a class of models that will later be efficiently fitted and then go through the diagnostic checking phase (Box and Jenkins, 2008).

There are rigorous ways to estimate the parameters of autoregressive models such as the method of least squares and Yule-Walker estimates. Moving average models can be estimated through the innovations method, see e.g. Brockwell and Davis (1991). The estimates of moving average models and the mixed models can also be obtained graphically or through iterative estimation procedures such as non-linear minimization (see e.g. Box and Jenkins, 2008).

Time series models should be able to describe the dependence among the observations, see e.g. Li (2004). It is a well-discussed issue that in time series model criticism, the residuals obtained from fitting a potential model to the observed time series play a vital role and can be used to detect departures from the underlying assumptions, (Box and Jenkins, 2008; Li, 2004).

In particular, if the model is a good fit to the observed series then the residuals should behave somewhat like a white noise process. So, taking into account the effect of estimation, the residuals obtained from a good fit should be approximately uncorrelated. While looking at the significance of residual autocorrelations, one approach is to test the significance of each individual residual autocorrelation which seems to be quite cumbersome. Another approach is to have some portmanteau test capable of testing the significance of the first, say m, residual autocorrelations (Box and Jenkins, 2008; Li, 2004), an approach we now describe.

2. Diagnostic Tests

The third stage of diagnostic checking process (Box and Jenkins, 2008) provides a practitioner an opportunity to test the model before using it for forecasting. This stage not only checks the fitted model for inadequacies but can also suggest improvements in the fitted model in the next iteration of this model building procedure. In this section we will do a literature review of the available diagnostic tests for fitted time series models.

In a time series context, if the fitted model is good then it should be able to explain the dependence pattern among successive observations. In other words, all the dependence in terms of autocorrelations and partial autocorrelations of the data generating process (DGP) should be explained by the fitted model so there should be no significant autocorrelation and partial autocorrelation in successive terms of the residuals.

In practice the most popular way for diagnostic checking a time series model is the portmanteau test, which tests whether any of a group of the first *m* autocorrelations $(\hat{r}_1,....,\hat{r}_m)$ of a time series are significantly different from zero. This type of test was first suggested by Box and Pierce (1970), in which they studied the distribution of residual autocorrelations in ARIMA processes. Based on the autocorrelations of the residuals obtained by fitting an ARMA(*p*, *q*) model to y_t, they suggested the following portmanteau test

$$Q_m = n\sum_{k=1}^{m}\hat{r}_k^2, \qquad (1)$$

where $\hat{r}_k$ is the residual autocorrelation at lag k. They suggested that $Q_m \sim \chi^2_{m-p-q}$, for moderate values of *m* and the fitted model is adequate, under the following conditions:

1. $\Psi_j \le O(n^{-1/2})$ for $j \ge m-p$, and
2. $\frac{m}{n} = O(n^{-1/2})$,

where Ψ_j are the weights in the MA(∞) representation.

Since Box and Pierce (1970) paper, the portmanteau test has become the vital part of time series diagnostic checking. Several modifications and versions of Box and Pierce (1970) has been suggested in the literature, see e.g. Ljung and Box (1978), McLeod and Li (1983), Monti (1994), Katayama (2008), Katayama (2009). These tests are capable of testing the significance of autocorrelation (partial autocorrelation) up to a finite number of lags. Chand et al. (2012) has used the portmanteau tests to criticize the fitted models.

In the discussion of Prothero and Wallis (1976), Chatfield has mentioned the poor power properties of Q_m and has recommended focusing on residual autocorrelations at the first few lags and seasonal lags. Similar suggestions are also made by Davies et al. (1977). In the same discussion on the Prothero and Wallis paper, Chatfield and Newbold also pointed out the poor approximation of the finite-sample distribution of Q_m. Prothero and Wallis (1976), in their reply to this discussion, suggested the use of the correction factor $(n + 2) / (n–k)$ to Q_m. However, this correction factor may inflate the variance of the

resulting statistic relative to that of the asymptotic χ^2_{m-p-q} distribution (see e.g. Davies et al., 1977, Ansley and Newbold, 1979).

An important point to note is that the statistic Q_m has been developed assuming the normality of the white noise process. As the results of Anderson and Walker (1964) suggest the asymptotic normality of the autocorrelation of a stochastic process is independent of the normality of the stochastic process and only depends on the assumption of finite variance, so the portmanteau test is expected to be insensitive to the normality assumption.

Ljung and Box (1978) suggested the use of the modified statistic

$$Q^*_m = n(n+2)\sum_{k=1}^{p}\frac{\hat{r}_k^2}{n-k}. \qquad (2)$$

They have shown that the modified portmanteau statistic Q^*_m has a finite sample distribution which is much closer to χ^2_{m-p-q}. Their results also show that Q^*_m is insensitive to the normality assumption of the error term, ε_t. As pointed out by many researchers e.g. Davies et al. (1977), Ansley and Newbold (1979), the true significance levels of Q_m tends to be much lower than predicted by the asymptotic theory and though the mean of Q^*_m is much closer to the asymptotic distribution, this corrected version of the portmanteau test has an inflated variance. But Ljung and Box (1978) pointed out that approximate expression of variance given by Davies et al. (1977) overestimates the variance of Q^*_m.

Several modifications have been suggested in Box and Pierce (1970) test and many useful versions of portmanteau statistic have been reported in literature, for example, Ljung and Box (1978), McLeod and Li (1983), Monti (1994), Pe˜na and Rodr´iguez (2002), Chand and Kamal (2006), Katayama (2008), and Katayama (2009).

Frequently in the literature larger values of *m* have been used in Q_m and Q^*_m, and the most commonly suggested value is *m* = 20 (see e.g. Davies et al., 1977, Ljung and Box, 1978). Ljung (1986) suggests the use of smaller values of *m* and has shown that for small values of *m*, Q^*_m has an approximate $a\chi^2_b$ distribution, where *a* and *b* are constants to be determined.

Ljung and Box (1978) also studied the empirical significance levels and empirical powers of Q^*_m for various choices of m and showed that the empirical significance levels for an AR(1) process are close to the nominal level for small choices of *m*, for example when *m* = 10 or 20, in all the cases except when the AR parameter is close to the stationarity region. This is a very challenging scenario for the χ^2 approximation. We will look at this issue in our future work. Ljung and Box (1978) also showed that approximating asymptotic distribution of $Q_m \sim \chi^2_v$, where $v = E(Q_m)$ results in performance of Q_m similar to that of $Q^*_m \sim \chi^2_{m-p-q}$.

As the partial autocorrelation function is an important tool in determining the order of an autoregressive process. Monti (1994) suggested a portmanteau test, following the idea of Ljung and Box (1978), given as:

$$Q_m^*(\hat{\omega}) = n(n+2)\sum_{k=1}^{m}\frac{\hat{\omega}_k^2}{n-k}, \qquad (3)$$

where $\hat{\omega}_k$ is the residual partial autocorrelation at lag *k*. She showed that $Q_m^*(\hat{\omega})$, analogously to Q_m^*, has an asymptotic null distribution χ^2_{m-p-q} and that $Q_m^*(\hat{\omega})$ is more powerful than Q_m^* especially when the order of the moving average component is understated.

As we have discussed earlier, the asymptotic distribution of Q_m and Q_m^* is questioned by several authors in the literature. Though small values of *m* solve this problem in some situations, it does not work in all cases, for example when the process is nearly stationary, see Ljung (1986). In a very recent paper, Katayama (2008) has suggested a bias corrected version

$$Q_m^{**} = Q_m^* - B_{m,n}^*$$

where

$$B_{m,n}^* = \hat{\mathbf{r}}^{\mathrm{T}}\mathbf{V}\mathbf{D}\mathbf{V}\hat{\mathbf{r}}, \hat{\mathbf{r}} = (\hat{r}_1,...,\hat{r}_m)^T, \mathbf{D} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \mathbf{V} = \mathrm{diag}\left(\sqrt{n(n+2)/(n-1)},...,\sqrt{n(n+2)/(n-m)}\right)$$

and $\boldsymbol{X}$ is an (m x (p + q)) matrix partitioned into p and q columns, see McLeod (1978) for details. Katayama (2008) showed the importance of this correction term especially for small values of *m* and when the roots of the ARMA(*p*,*q*) process lie near the stationarity region.

In practice, the optimal choice of *m* is difficult as the use of the χ^2_{m-p-q} approximation and diagnostic checking require large values of *m* which results in less power and unstable size of test, as noticed by Ljung (1986), Katayama (2008). Katayama (2009) suggested a multiple portmanteau test to overcome this problem. His suggested test is based on several portmanteau tests for a range of small to medium values of m. He showed using some numerical examples that his suggestion leads to a superior test. He also discussed the linkage between his suggested multiple test and the test due to Pena and Rodriguez (2002). He suggested a method based on some iterative procedure to approximate joint distribution of the multiple test as the computation of the distribution is very hard due to correlated elements.

For the past few decades with the advent of high-speed computers, the interest of researchers have been focused on nonlinear models. It has been pointed out by several researchers that the Box-Pierce type tests fail to show good power against nonlinear models (see e.g. Escanciano, 2006; Pena and Rodriguez, 2002). McLeod and Li (1983) used the sample autocorrelation of the squared residuals to test for linearity against the nonlinearity and showed its good power against departures from linearity.

Escanciano (2007) proposed diagnostic tests based on the CvM test using the weights suggested by Bierens (1982), given by

$$C_vM_{\exp P} = \frac{1}{n\hat{\sigma}^2}\sum_{t=1}^{n}\sum_{s=1}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_s \exp\left(-\frac{1}{2}\left|\mathbf{I}_{t-1,P}-\mathbf{I}_{s-1,P}\right|^2\right), \quad (4)$$

where $\hat{\sigma}^2 = \sum_{t=1}^{n}\hat{\varepsilon}_t^2/n-1$ is the variance of residuals and

$$\mathbf{I}_{t-1,P} = (y_{t-1}, y_{t-2}, ..., y_{t-P}) \quad (5)$$

is the information set at time t – 1 and dimension *P*. It can be noticed that the distance $\left|\mathbf{I}_{t-1,P}-\mathbf{I}_{s-1,P}\right|^2$ increases very fast with *P* which results in weights being near 0 when P is relatively large. We have considered the CvM statistic with this weight scheme in our study as it has shown good power properties reported in Escanciano (2006).

3. Methodology

We now consider various versions of the statistics defined in (1), (2), (3) and (4). We compare empirical size and power of these tests against various linear and non-linear classes of models. Mainly we compare the dynamic and fixed design bootstrap methods but we also look at the usefulness of transformed residuals in bootstrap methods.

3.1 Bootstrap Methods

For time series data, the dependence structure of the data generating process (DGP) makes it difficult to apply the bootstrap methods. In general, there are two main bootstrap methods that are used in time series i.e. model-based bootstrap methods and block-resampling bootstrap methods. Generally, the model-based bootstrap methods are called resampling-residuals bootstrap methods.

In block bootstrapping, we divide the sample into overlapping or non-overlapping blocks of a certain length. The performance of block bootstrap methods much depends on block length. Under the stationarity condition each block should have the same joint probability distribution. In our study we consider only the model-based bootstrapping, as model based bootstrap methods tend to be more accurate than block bootstrap methods (Lahiri, 2003) and also as our objective is to compare two model-based bootstrap methods, namely dynamic bootstrap and fixed design bootstrap.

Suppose we have a sample time series $\{y_t\}_{t=1}^{n}$ generated by a DGP defined by

$$y_t = f(\mathbf{I}_{t-1,P}, \theta) + \varepsilon_t, \quad (6)$$

where $\mathbf{I}_{t-1,P}$ is the information set defined earlier in (5) and θ is the vector of model parameters. Suppose the fitted model is

$$\hat{y}_t = f(\mathbf{I}_{t-1,P}, \hat{\theta}), \ t = P, P+1, ...$$

where $\hat{\theta}$ is the estimate of θ. Thus the residuals are

$$\hat{\varepsilon}_t = y_t - \hat{y}_t, \quad (7)$$

We assume that initial data $y_{t-P}, ..., y_0$ are available.

Semi-parametric time series bootstrap methods

Under the assumption that the DGP given in (6) is the true model for the given sample time series, the residuals given in (7) will serve the purpose of an IID sample. The following approaches are used in semi-parametric time series bootstrap methods.

Dynamic bootstrap. If the error terms, $\varepsilon_t's,$ in our DGP are IID, with common variance $\sigma^2,$ then we can generally make very accurate inferences by using the dynamic bootstrap (DB) (MacKinnon, 2006). This method requires the IID assumption of the error term and only mild conditions on its distribution. The DB is defined as:

$$y_t^* = f\left(\mathbf{I}_{t-1,P}^*, \hat{\theta}\right) + \varepsilon_t^* \qquad \text{for } t = 1,2,.....,n, \tag{8}$$

where $\mathbf{I}_{t-1,P}^* = (y_{t-1}^*,...,y_{t-P}^*)$ is the dynamic bootstrap of the information set defined in (5) and ε_t^* is selected at random with replacement from the vector of the residuals $(\hat{\varepsilon}_1, \hat{\varepsilon}_2,..., \hat{\varepsilon}_n).$

Dynamic wild bootstrap The dynamic wild bootstrap (DWB) is a simple modification of the dynamic bootstrap. The only difference is to resample the rescaled residuals instead of fitted residuals. These rescaled residuals are usually named as wild bootstrap. Various rescaling schemes have been suggested in the literature, see e.g. Liu (1988) or Stute et al. (1998). The DWB is defined as:

$$y_t^o = f\left(\mathbf{I}_{t-1,P}^{\mathbf{o}}, \hat{\theta}\right) + \varepsilon_t^o \qquad \text{for } t = 1,2,.....,n, \tag{9}$$

where $\mathbf{I}_{t-1,P}^{\mathbf{o}} = (y_{t-1}^o,...,y_{t-P}^o)$ is the DWB of the information set defined in (5) and $\varepsilon_t^o = \hat{\varepsilon}_t.v_t,$ such that the sequence v_t is IID with zero mean, unit variance and finite fourth moment. Liu (1988) has suggested the following v_t for transforming the IID residuals to wild residuals,

$$v_t = \begin{cases} -1 & with\ p = \dfrac{1}{2} \\ +1 & with\ p = \dfrac{1}{2} \end{cases}$$

Stute et al. (1998) has suggested the following as v_t sequence,

$$v_t = \begin{cases} \dfrac{1-\sqrt{5}}{2} & with\ p = \dfrac{1+\sqrt{5}}{2\sqrt{5}} \\ \dfrac{1+\sqrt{5}}{2} & with\ p = 1 - \dfrac{1+\sqrt{5}}{2\sqrt{5}} \end{cases}$$

The fixed design wild bootstrap In fixed design wild bootstrap (FWB), the bootstrap sample is generated from the fixed design $\mathbf{I}_{t-1,P}$. Moreover, the fitted residuals are transformed to wild residuals using the suggested transformations (see Liu, 1988 and Stute et al., 1998). The FWB is defined as:

$$y_t^\Diamond = f\left(\mathbf{I}_{t-1,P}, \hat{\theta}\right) + \varepsilon_t^\Diamond \qquad \text{for } t = 1,2,.....,n, \tag{10}$$

where $\varepsilon_t^\Diamond = \hat{\varepsilon}_t.v_t$ and v_t is as defined above.

4. Model Estimation

In this section, we describe the estimation methods used in our study. In our numerical results showed in Section 5. Given an AR(p) model

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \mathrm{L} + \alpha_p y_{t-p} + \varepsilon_t$$

we estimate the AR(p) model under various bootstrap designs discussed earlier in Section 3.1. The ordinary least squares (OLS) estimates (Gonçalves and Kilian, 2004) of $\alpha = (\alpha_1, ..., \alpha_p)$ are obtained as below:

$$\hat{\alpha}^* = \left(n^{-1} \sum_{t=1}^{n} \mathbf{I}^*_{t-1,P} \mathbf{I}^{*T}_{t-1,P} \right)^{-1} n^{-1} \sum_{t=1}^{n} \mathbf{I}^*_{t-1,P} y^*_t,$$

$$\hat{\alpha}^o = \left(n^{-1} \sum_{t=1}^{n} \mathbf{I}^o_{t-1,P} \mathbf{I}^{oT}_{t-1,P} \right)^{-1} n^{-1} \sum_{t=1}^{n} \mathbf{I}^o_{t-1,P} y^o_t,$$

$$\hat{\alpha}^\diamond = \left(n^{-1} \sum_{t=1}^{n} \mathbf{I}_{t-1,P} \mathbf{I}^{T}_{t-1,P} \right)^{-1} n^{-1} \sum_{t=1}^{n} \mathbf{I}_{t-1,P} y^\diamond_t,$$

where $\mathbf{I}_{t-1,P}$ is the information set defined in (5) while $\mathbf{I}^*_{t-1,P}$ and $\mathbf{I}^o_{t-1,P}$ are the information sets for DB and DWB respectively defined in Section 3.1. The y^*_t, y^o_t and $y^\diamond_t$ are defined in (8), (9) and (10).

In this paper, we mainly look at the power of the diagnostic tests. We use the bootstrap distributions under the semi-parametric bootstrap designs discussed in Section 3.1. We also look at empirical power of test against various alternative models.

4.1 Algorithms

In this section, we give the algorithms for the Monte Carlo method used to compute the empirical size and power of the diagnostic tests defined in Section2. For each Monte Carlo run, a sample time series $\{y_t\}_{t=1}^n$ is simulated under the model M. For the sake of convenience, we denote the statistic of interest as T. For the computation of power M is the alternative model. We estimate the null model for the simulated sample time series and T is calculated from the residuals, $\hat{\varepsilon}_t$.

Algorithm 1: Bootstrap sampling procedure

Step 1	Generate bootstrap sample y^*_t from resamples of $\hat{\varepsilon}_t$, say $\hat{\varepsilon}^*_t$.
Step 2	Fit the null model to the bootstrap sample y^*_t and obtain residuals as $\hat{\varepsilon}^*_t = y^*_t - \hat{y}^*_t$, where $\hat{y}^*_t$ is the fitted series.
Step 3	Using the residuals, $\hat{\varepsilon}^*_t$, calculate test-statistic T, say, T^*.
Step 4	Step 4 Repeat Step 1-3 for each of the B bootstrap samples.

Algorithm 1 gives the bootstrap procedure used in our numerical study. From this algorithm, we obtain the bootstrap approximation of the distribution of the test. We will use this algorithm to compute power in the following algorithms of our simulation study consisting of N Monte Carlo runs.

The power of a test is the probability of rejecting a false null hypothesis. For empirical power, as mentioned earlier, the sample is generated under the alternative model. Algorithm 2 state the Monte Carlo procedure we use to determine the power of test.

Algorithm 2: Computation of empirical power

Step 1	Calculate $100(1-\alpha)th$ percentile, say $T^*_{1-\alpha}$, of the bootstrap distribution of T^* obtained using Algorithm 1.
Step 2	Reject null model if $T \geq T^*_{1-\alpha}$ otherwise accept it.
Step 3	Repeat Step 1-2 for each of the N Monte Carlo runs.
Step 4	Empirical power, $1-\hat{\beta}$, is determined as below, $$1-\hat{\beta} = \frac{\#(T \geq T_{1-\alpha})}{N}$$

In the next section, we look at different examples and compute the power of the diagnostic tests. Now we give definitions of some nonlinear models which we will study as alternative models in empirical power study of portmanteau tests.

Exponential Autoregressive model

An exponential autoregressive model, EXPAR(p), is defined as

$$y_t = \sum_{i=1}^{p}[\alpha_i + \pi_i \exp(-\gamma y_{t-1}^2)] + \varepsilon_t$$

For detailed discussion see e.g. Tong (1990, p.108).

Threshold Autoregressive model

The threshold autoregressive, TAR(p), model is defined as

$$y_t = \begin{cases} \sum_{i=1}^{p} \alpha_i^{(1)} y_{t-i} + \varepsilon_t & \text{if } y_{t-i} < r \\ \sum_{i=1}^{p} \alpha_i^{(2)} y_{t-i} + \varepsilon_t & \text{if } y_{t-i} \geq r \end{cases}$$

where r is called the threshold, below r the AR parameters are $\alpha_i^{(1)}$ and above r these are $\alpha_i^{(2)}$ (see e.g. Chatfield, 2004, p.200). Threshold models were developed and introduced by Tong and Lim (1980) which are basically piecewise linear AR models. For more discussion on bilinear models see also Tang and Mohler (1988) and references therein.

5. Results and Discussion

In this section, we look at some numerical examples to compare the empirical power of the goodness of fit tests under various semi-parametric bootstrap designs discussed in Section 3.1. We present and compare the power against linear and non-linear alternative class of models under a linear null model. Empirical power results are obtained using Algorithm 2 consisting of 1000 Monte Carlo runs of 200 bootstrap samples. Each bootstrap sample is of size n = 100.

5.1 Linear Alternatives

Mixed ARMA models are the most commonly used models in applications. In this section we compare the power of the tests against several versions of ARMA(2, 2). In this example, we simulate the series for the alternative model, ARMA(2, 2) process, given below:

$$y_t = 1.05 + 1.41y_{t-1} - 0.77y_{t-2} + 0.33k\varepsilon_{t-1} + 0.21k\varepsilon_{t-2} + \varepsilon_t,$$

where $\varepsilon_t \sim N(0,1)$. We fit an AR(2) model to this sample and the power results in the following table of the percentage of time we rejected the null model. Importantly, note that we consider various values of *k* ranging from 0 to 2. It can be noticed that choice *k* = 0 corresponds to an AR(2) process, so we expect very low power in this case, actually as low as the level of significance. On the other hand, as the value of k increases, the MA component in an ARMA process increases in absolute value and this should result in a higher power, reaching a maximum of 100%, for some value of *k*.

	$k = 0$			$k = 0$			$k = 0$		
	DB	DWB	FWB	DB	DWB	FWB	DB	DWB	FWB
C_vM_{exp3}	6.1	6.2	5.4	10.7	10.1	10.6	20.0	20.5	20.6
C_vM_{exp5}	4.8	4.4	4.7	7.5	7.7	8.5	15.1	15.0	16.7
C_vM_{exp7}	4.9	4.9	5.3	10.0	9.9	10.6	11.7	13.0	14.9
Q_5	5.2	5.1	2.0	42.7	43.3	26.4	99.2	99.3	97.8
Q_{10}	5.6	5.8	2.3	33.1	34.1	21.2	96.1	96.7	93.1
Q_{25}	5.2	4.9	1.5	29.4	27.9	20.4	90.9	91.7	85.9
Q_5^*	5.3	5.3	2.0	42.0	42.2	25.9	99.1	99.1	97.8
Q_{10}^*	5.9	6.0	2.4	32.3	32.9	20.7	95.6	96.1	92.5
Q_{25}^*	5.7	5.0	2.3	28.0	26.9	18.9	88.9	88.1	82.4
$Q_5^*(\hat{\omega})$	4.8	5.0	1.2	47.7	47.1	30.2	99.5	99.6	98.4
$Q_{10}^*(\hat{\omega})$	4.6	6.0	2.2	33.6	33.7	21.3	97.9	98.3	95.6
$Q_{25}^*(\hat{\omega})$	4.4	4.7	2.7	26.3	26.2	19.1	91.7	91.1	87.1

Table 1: Power (in %), based on 1000 Monte Carlo runs of 200 bootstrap samples of size 100 for AR(2), against ARMA(2,2), $y_t = 1.05 + 1.41y_{t-1} - 0.77y_{t-2} + 0.33k\varepsilon_{t-1} + 0.21k\varepsilon_{t-2} + \varepsilon_t$.

Table 1 gives the results for empirical power of the goodness-of-fit tests. It can be very clearly noticed that CvM has less power while portmanteau tests have better power against this linear class of alternatives. Our results confirm the results reported in the literature, see e.g Hong and Lee (2003); Escanciano (2006). Though we have provided the power results for both of dynamic and fixed design bootstrap methods, we discuss the results for dynamic bootstrap method only, as dynamic bootstrapping provides the best approximation to the asymptotic distribution especially for the portmanteau tests, see e.g. Chand (2013).

We can see from these results as we increase the value of *k*, in general, the power for each of the goodness-of-fit tests increases but the increase that for $C_vM_{\exp P}$ is not exponential and it attains a maximum power around 20% even for k = 2. In contrast to this, the portmanteau tests show an exponential increase in power with an increase in k and reaches nearly to maximum power of 100%.

Moreover, it can also be seen that as the value of m increases for the portmanteau tests, these tests become generally less powerful. This result is well known and reported in the literature, see e.g. Hong and Lee (2003), Katayama (2009). The same kind of behaviour can be seen for $C_vM_{\exp P}$ test and it also shows a decrease in power for larger values of P, this is also reported in Escanciano (2006).

5.2 Non Linear Alternatives

In this section, we look at the empirical power of the goodness-of-fit tests against some popular non-linear alternatives. We consider several versions of non linear EXPAR(2) and TAR(2) models. It has been reported in the literature that the portmanteau tests, we are studying, have poor power against non-linear alternatives especially for TAR models (Escanciano, 2006). We will use the same choices of P and m as we have used in previous section of power against linear alternatives i.e. P = 3, 5, 7 for $CvM_{\exp,P}$ test and m = 5, 10, 25 for residual autocorrelations based portmanteau tests.

First, we take an EXPAR(2) model, defined as

$$y_t = (0.138 + k(0.316 + 0.316 + 0.982y_{t-1})e^{(-3.89y_{t-1}^2)})y_{t-1} - (0.437$$

$$+ k(0.659 + 1.260y_{t-1})e^{(-3.89y_{t-1}^2)})y_{t-2} + 0.2\varepsilon_t,$$

where $\varepsilon_t \sim N(0,1)$. The empirical power of diagnostic tests is computed using Algorithm 2.

	$k = 0.8$			$k = 1.0$			$k = 2.0$		
	DB	DWB	FWB	DB	DWB	FWB	DB	DWB	FWB
$C_v M_{\exp 3}$	29.6	30.0	21.2	69.5	70.2	61.0	100	100	100
$C_v M_{\exp 5}$	26.8	27.9	21.8	67.1	67.2	61.2	100	100	100
$C_v M_{\exp 7}$	21.8	22.3	18.3	62.2	63.1	56.0	100	100	100
Q_5	8.7	8.5	3.5	11.2	11.7	5.2	42.3	42.2	23.1
Q_{10}	5.7	6.2	2.3	8.6	7.5	4.4	29.5	29.0	18.0
Q_{25}	5.8	4.9	2.7	8.0	7.5	4.3	31.1	30.7	22.3
Q_5^*	8.8	8.7	3.5	11.1	11.9	5.3	43.0	43.2	24.0
Q_{10}^*	5.8	5.9	2.1	8.4	7.5	4.4	29.6	28.7	18.0
Q_{25}^*	6.1	5.2	3.0	8.0	7.7	4.1	29.7	30.0	21.4
$Q_5^*(\hat{\omega})$	8.6	8.3	2.8	11.9	11.9	5.6	40.6	39.4	22.6
$Q_{10}^*(\hat{\omega})$	5.6	6.1	2.6	8.6	7.8	3.3	28.7	27.9	16.9
$Q_{25}^*(\hat{\omega})$	4.9	5.0	3.0	7.7	7.3	4.5	28.5	28.9	22.6

Table 2: Power (in %), based on 1000 Monte Carlo runs of 200 bootstrap samples of size 100 for AR(2) against EXPAR(2).

Table 2 reports the empirical power of the diagnostic tests. The situation looks quite opposite to the linear case in the previous section. As we can see, $k = 0$ will correspond to an AR(2) process and with an increase in value of k, the non-linear component in the model will become dominant.

The results in Table 2 suggest that residual autocorrelations based portmanteau tests have low power against this class of non-linear alternatives while $C_v M_{\exp P}$ is showing good power in this case. As it can be seen that $C_v M_{\exp P}$ power increases exponentially with an increase in k and attains the maximum power 100% at $k = 2$ while power for the portmanteau tests can reach around 43%. These results confirm our earlier findings that power decreases for larger values of P and m.

Now, we move to threshold autoregressive model, another class of non-linear models. Theory suggests that TAR models are more challenging than EXPAR models for the diagnostic tests. We consider the following TAR(2) model

$$y_t = \begin{cases} (1.435 - 0.815k) + (1.385 - 0.135k) y_{t-1} \\ \qquad + (-0.835 + 0.405k) y_{t-2} + \varepsilon_t \quad for \quad y_{t-2} \le 3.25 \\ (1.435 - 0.815k) + (1.385 - 0.135k) y_{t-1} \\ \qquad + (-0.835 + 0.405k) y_{t-2} + \varepsilon_t \quad for \quad y_{t-2} \le 3.25 \end{cases}$$

where $\varepsilon_t \sim N(0,1)$. We can see by controlling the value of k, we can control the amount of nonlinearity in the model. The lower value of k corresponds to low levels of nonlinearity while larger values of k will result in a highly nonlinear model. We use a range of values

of k where the model does not blow up. We use the same Algorithm 2 to compute the empirical power.

	$k = 0.8$			$k = 1.0$			$k = 1.5$		
	DB	DWB	FWB	DB	DWB	FWB	DB	DWB	FWB
$C_vM_{\exp,3}$	43.8	42.7	40.2	49.1	49.0	45.6	48.0	49.1	43.5
$C_vM_{\exp,5}$	24.7	23.9	22.2	29.9	29.5	26.0	32.4	32.0	29.0
$C_vM_{\exp,7}$	14.0	13.7	12.1	14.9	14.1	12.5	18.5	18.9	16.7
Q_5	5.8	5.4	1.6	4.8	4.8	1.3	6.0	5.2	1.6
Q_{10}	5.3	6.6	2.5	5.4	5.6	1.5	4.5	4.4	1.5
Q_{25}	6.8	6.5	3.4	6.8	6.8	3.0	5.0	4.5	2.0
Q_5^*	5.7	5.4	1.6	4.9	4.7	1.3	6.0	5.1	1.5
Q_{10}^*	5.4	6.5	2.5	5.2	5.6	1.7	4.4	4.4	1.6
Q_{25}^*	6.7	6.8	3.4	6.3	7.0	7.2	4.8	4.8	1.9
$Q_5^*(\hat{\omega})$	6.1	5.9	2.3	6.0	6.3	1.1	5.8	5.3	1.9
$Q_{10}^*(\hat{\omega})$	5.7	5.9	1.7	5.9	5.3	2.2	5.2	5.9	1.0
$Q_{25}^*(\hat{\omega})$	7.2	6.5	3.6	6.8	7.2	4.1	5.7	5.5	3.2

Table 3: Power (in %), based on 1000 Monte Carlo runs of 200 bootstrap samples of size 100 for AR(2), against TAR(2).

Table 3 reports the empirical power of the diagnostic tests for AR(2) against TAR(2) models. These results generally confirm the known fact that threshold models are challenging for the goodness of fit tests. The residual autocorrelations based portmanteau tests show very low power against the TAR model. Though $CvM_{\exp,P}$ is showing better power results especially for smaller choice of P, i.e. $P = 3$, it still cannot achieve the same high power as it did against EXPAR(2) models.

Importantly, it should be noted that choice of P and m is very crucial and the power results may improve for some smaller values of P and m. Noting the result reported in Escanciano (2006), where $C_vM_{\exp P}$ has achieved power of81% against TAR(1) model, we tried smaller values of P, i.e. $P = 1, 2$. For $P = 1$, power for $CvM_{\exp,P}$ even further decreases to around 20% while for $P = 2$, it shows an improvement and power rises to 60%.

6. Conclusion

Portmanteau tests are powerful against the linear alternatives while the *CvM* statistic has shown more power against non-linear alternatives. The choice of *m* for portmanteau tests and *P* for $CvM_{\exp,P}$ test is important. Our results suggest that approximation of the asymptotic distribution and power of these goodness-of-fit tests highly depends on the choice of these parameters, *P* and *m*.

References

1. Anderson, T. and Walker, A. (1964). On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process. The Annals of Mathematical Statistics, 35(3): 1296-1303.
2. Ansley, C. and Newbold, P. (1979). On the finite sample distribution of residual autocorrelations in autoregressive-moving average models. Biometrika, 66(3): 547-553.
3. Bierens, H. (1982). Consistent model specification tests. Journal of Econometrics, 20(1): 105-134.
4. Box, G. and Jenkins, G. (2008). Time Series Analysis, Forecasting and Control. John Wiley & Sons: New Jersey, USA.
5. Box, G. and Pierce, D. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association, 65(332): 1509-1526.
6. Brockwell, P. and Davis, R. (1991). Time Series: Theory and Methods. Springer-Verlag: New York, USA.
7. Chand, S. and Kamal, S. (2006). A comparative study of portmanteau tests for univariate time series models. Pakistan Journal of Statistics and Operation Research, 2(2).
8. Chand, S., Kamal, S., and Ali, I. (2012). Modelling and volatility analysis of share prices using ARCH and GARCH models. World Applied Sciences Journal, 19(1): 77–82.
9. Chand, S. (2011). Goodness of fit tests and lasso variable selection in time series analysis. *PhD thesis, University of Nottingham*.
10. Chand, S. (2013). Bootstrapping Time Series Goodness of Fit Tests. *Submitted paper.*
11. Chatfield, C. (2004). The Analysis of Time Series: An Introduction. Chapman & Hall/CRC: London, UK.
12. Davies, N., Triggs, C., and Newbold, P. (1977). Significance levels of the Box-Pierce portmanteau statistic in finite samples. Biometrika, 64(3): 517-522.
13. Escanciano, J. (2006). Goodness-of-fit tests for linear and nonlinear time series models. Journal of the American Statistical Association, 101(474):531-541.
14. Escanciano, J. (2007). Model checks using residual marked empirical process. Statistica Sinica, 17(1): 115-138.

15. Gonçalves, S. and Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. Journal of Econometrics, 123(1): 89-120.

16. Hong, Y. and Lee, T. (2003). Diagnostic checking for the adequacy of nonlinear time series models. Econometric Theory, 19(6): 1065-1121.

17. Horowitz, J., Lobato, I., Nankervis, J., and Savin, N. (2006).Bootstrapping the Box-Pierce Q test: A robust test of uncorrelatedness. Journal of Econometrics, 133(2): 841-862.

18. Katayama, N. (2008). An improvement of the portmanteau statistic. Journal of Time Series Analysis, 29(2): 359-370.

19. Katayama, N. (2009). On multiple portmanteau tests. Journal of Time Series Analysis, 30(5): 487-504.

20. Lahiri, S. (2003). Resampling Methods for Dependent Data. Springer-Verlag: New York, USA.

21. Li, W. (2004). Diagnostic Checks in Time Series. Chapman & Hall/CRC: New York, USA.

22. Liu, R. (1988). Bootstrap procedures under some non-IID models. The Annals of Statistics, 16(4): 1696-1708.

23. Ljung, G. (1986). Diagnostic testing of univariate time series models. Biometrika, 73(3):725-730.

24. Ljung, G. and Box, G. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2): 297-303.

25. MacKinnon, J. (2006). Bootstrap methods in econometrics. The Economic Record, 82(1): S2-S18.

26. Mainassara, B. et al. (2009). Multivariate portmanteau test for structural VARMA models with uncorrelated but non-independent error terms. MPRA Paper.

27. McLeod, A. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. Journal of the Royal Statistical Society. Series B (Methodological), 40(3): 296-302.

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

29. Monti, A. (1994). A proposal for a residual autocorrelation test in linear models. Biometrika, 81(4): 776-780.

30. Pena, D. and Rodriguez, J. (2002). A powerful portmanteau test of lack of fit for time series. Journal of the American Statistical Association, 97(458): 601-611.

31. Prothero, D. and Wallis, K. (1976). Modelling macroeconomic time series. Journal of the Royal Statistical Society. Series A (General), 139(4): 468-500.

32. Stute, W., Manteiga, W., and Quindimil, M. (1998). Bootstrap approximations in model checks for regression. Journal of the American Statistical Association, 93(441): 141-149.

33. Tang, Z. and Mohler, R. (1988). Bilinear Time Series: Theory and Application. Springer-Verlag: New York, USA.

34. Tong, H. (1990). Nonlinear Time Series: a Dynamical System Approach. Oxford University Press: Oxford, UK.

35. Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. Journal of the Royal Statistical Society. Series B (Methodological), 42(3): 245-292.