Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 1731–1738
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Improved Peña–Rodriguez portmanteau test

Jen-Wen Lin, A.Ian McLeod*

Department of Statistical and Actuarial Sciences, The University of Western Ontario, London, Ont., Canada N6A 5B7

Received 27 February 2006; received in revised form 9 June 2006; accepted 10 June 2006
Available online 10 July 2006

Abstract

Several problems with the diagnostic check suggested by Peña and Rodriguez [2002. A powerful portmanteau test of lack of fit for time series. J. Amer. Statist. Assoc. 97, 601–610.] are noted and an improved Monte-Carlo version of this test is suggested. It is shown that quite often the test statistic recommended by Peña and Rodriguez [2002. A powerful portmanteau test of lack of fit for time series. J. Amer. Statist. Assoc. 97, 601–610.] may not exist and their asymptotic distribution of the test does not agree with the suggested gamma approximation very well if the number of lags used by the test is small. It is shown that the convergence of this test statistic to its asymptotic distribution may be quite slow when the series length is less than 1000 and so a Monte-Carlo test is recommended. Simulation experiments suggest the Monte-Carlo test is usually more powerful than the test given by Peña and Rodriguez [2002. A powerful portmanteau test of lack of fit for time series. J. Amer. Statist. Assoc. 97, 601–610.] and often much more powerful than the Ljung–Box portmanteau test. Two illustrative examples of enhanced diagnostic checking with the Monte-Carlo test are given.

Keywords: ARMA residual diagnostic test; Imhof distribution; Monte-Carlo test; Portmanteau diagnostic check

1. Introduction

Let $X_t, t = 1, 2, \ldots$ be a stationary and invertible ARMA (p, q) model (Box et al., 1994),

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) X_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right) a_t, \tag{1}$$

where B is the backshift operator on t, a_t is a sequence of independent and identical normal random variables with mean zero and variance σ_a^2. After fitting this model to a series of length n, the residual autocorrelations,

$$\hat{r}(k) = \sum_{t=k+1}^{n} \hat{a}_t \hat{a}_{t-k} / \sum_{t=1}^{n} \hat{a}_t^2, \quad k = 1, 2, \ldots, \tag{2}$$

where $\hat{a}_t$ denotes the fitted residuals, may be used for checking model adequacy. One of the most widely used model

* Corresponding author. Tel.: +1 519 6613611; fax: +1 519 6613813.
E-mail address: aimcleod@uwo.ca (A.Ian McLeod).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.010

diagnostic checks (Li, 2004) is the portmanteau test of Ljung and Box (1978),

$$Q_m = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\hat{r}(k)^2, \tag{3}$$

where, under the assumption of model adequacy, Q_m is approximately $\chi^2_{m-(p+q)}$ distributed.

A new portmanteau diagnostic test statistic, $\hat{D}_m = n\left(1 - \left|\hat{R}_m\right|^{1/m}\right)$, based on the determinant of the residual autocorrelation matrix,

$$\hat{R}_m = \begin{pmatrix} 1 & \hat{r}(1) & \cdots & \hat{r}(m) \\ \hat{r}(1) & 1 & \cdots & \hat{r}(m-1) \\ \vdots & \cdots & \ddots & \vdots \\ \hat{r}(m) & \cdots & \hat{r}(1) & 1 \end{pmatrix}, \tag{4}$$

was suggested by Peña and Rodriguez (2002). As noted by Peña and Rodriguez (2002), $\left|\hat{R}_m\right|$ is the estimated generalized variance of the residuals standardized by dividing by their standard deviation.

2. The D_m test and its limitations

Peña and Rodriguez (2002, Theorem 1) showed that if the model is correctly identified, $\hat{D}_m$ is asymptotically distributed as $\sum_{i=1}^{m}\lambda_i\chi^2_{1,i}$, where $\chi^2_{1,i}$ are independent chi-squared random variables with one degree of freedom, and λ_i are the eigenvalues of $\mathcal{Q}_m W_m$, where W_m is a diagonal matrix with the ith diagonal elements, $w_i = (m-i+1)/m$, $i = 1, 2, \ldots, m$ and $\mathcal{Q}_m$ is the asymptotic covariance matrix of the normalized residual autocorrelations $\sqrt{n}(\hat{r}(1), \ldots, \hat{r}(m))$ given in McLeod (1978, Eq. 15). As pointed out by Peña and Rodriguez (2002), this asymptotic distribution may be computed using the method of Imhof (1961). We have implemented the computation of this asymptotic distribution for the general ARMA(p, q) model in our R package `gvtest`. The cumulative distribution function for this asymptotic distribution may be denoted by $F(x; \lambda_1, \ldots, \lambda_m)$.

Peña and Rodriguez (2002) suggested evaluating $\hat{D}_m$ by a gamma approximation distribution to the asymptotic distribution. The approximation distribution is derived by equating the first two moments of a gamma distribution with those of the corresponding asymptotic distribution. The density function for this gamma approximation may be written $f_\gamma(x; \alpha, \beta) = \mathrm{e}^{-x/\beta}x^{\alpha-1}\beta^{-\alpha}/\Gamma(\alpha)$, where

$$\alpha = 3m\{(m+1) - 2(p+q)\}^2/[2\{2(m+1)(2m+1) - 12m(p+q)\}] \tag{5}$$

and

$$\beta = 3m\{(m+1) - 2(p+q)\}/\{2(m+1)(2m+1) - 12m(p+q)\}. \tag{6}$$

Peña and Rodriguez (2002) indicated that the approximation improves as m increases. In addition, it may be noted that if m is too small then it can happen that $\alpha \leqslant 0$ or $\beta \leqslant 0$ which is numerically infeasible. For example, if $p+q=3$ then we must have $m \geqslant 8$ to make $\alpha > 0$ and $\beta > 0$.

Peña and Rodriguez (2002) found that the empirical distribution of $\hat{D}_m$ did not agree very well with the gamma approximation so they suggested a modified statistic,

$$D_m = n - n\left|\ddot{R}_m\right|^{1/m}, \tag{7}$$

where $\ddot{R}_m$ denotes the residual autocorrelation matrix replacing $\hat{r}(k)^2$ with $\ddot{r}_k^2$, where

$$\ddot{r}(k)^2 = (n+2)(n-k)^{-1}\hat{r}(k)^2.$$

This is similar in spirit to the modification suggested by Ljung and Box (1978) to the original Box and Pierce portmanteau statistic (1970). Although, as shown by Peña and Rodriguez (2002), this approximation works well when $n \leqslant 100$ in first-order autoregressive models, it does not provide a good approximation to the asymptotic distribution for more

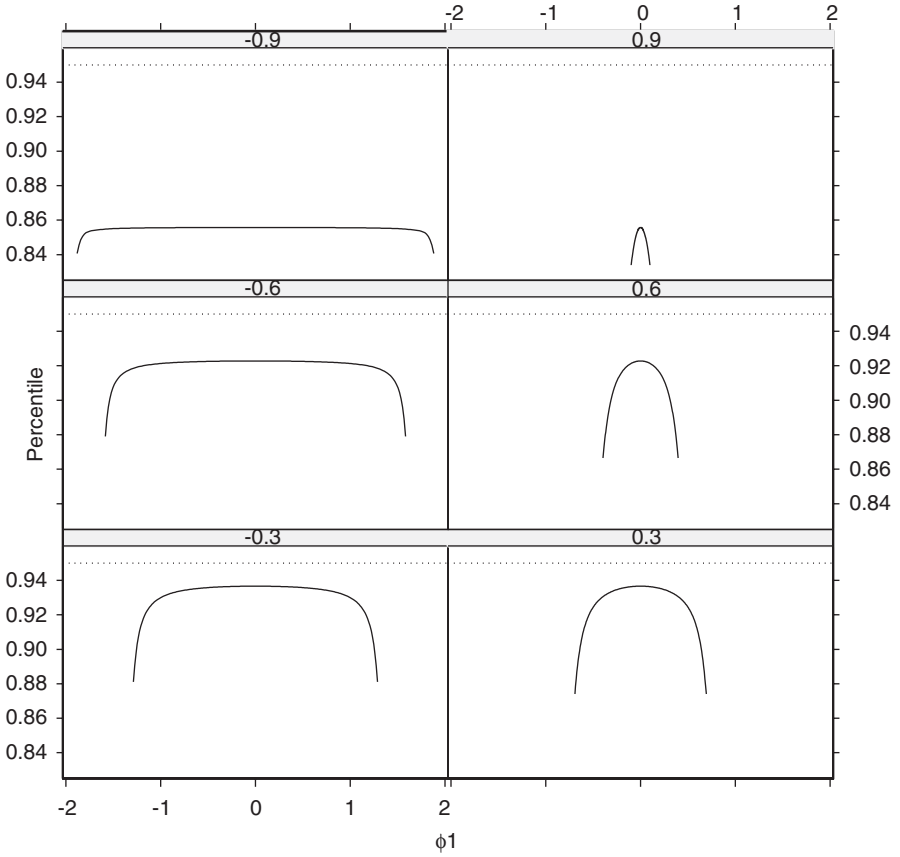

Fig. 1. Inaccuracy of the gamma approximation for AR(2) with $m = 10$. $F(Q_\gamma(0.05; \alpha, \beta); \lambda_1, \ldots, \lambda_1 0)$, where F is the cdf of asymptotic distribution of $\hat{D}_m$ and Q_γ is the quantile function for the gamma approximation, is plotted. The plot shows panels corresponding to $\phi_2 = \pm 0.3, \pm 0.6, \pm 0.9$. The parameter ϕ_1 extends over the whole admissible region corresponding to the ϕ_2. These plots demonstrate that the gamma approximation is not conservative asymptotically. When n is large enough, the test based on the gamma approximation will overstate the true significance level.

Table 1
P-values of the upper 5% quantiles of the gamma approximation for ARMA (1, 1) models with ϕ_1 and $\theta_1 = \pm 0.9, \pm 0.6, \pm 0.3$ evaluated by their asymptotic distributions

$\theta_1 \mid \phi_1$	−0.9	−0.6	−0.3	0.3	0.6	0.9
−09		0.105	0.091	0.083	0.085	0.109
−06	0.105		0.069	0.063	0.065	0.085
−03	0.091	0.692		0.060	0.063	0.083
03	0.083	0.063	0.060		0.069	0.091
06	0.085	0.065	0.063	0.069		0.105
09	0.108	0.085	0.083	0.091	0.105	

complicated models if the number of lags, m, is small. As shown in Fig. 1, the gamma approximation can distort the size of a 5% significance test relative to the asymptotic distribution for AR (2) models when m is 10. Similar distortions are found for MA (2) models or higher orders AR and MA models. The distortions for ARMA (1, 1) with $m = 10$ models were also investigated and the results were listed in Table 1. Moreover, this distortion would tend to make the tests based on the gamma approximation reject more often than they should. In other words, the test based on the gamma approximation is not conservative. So despite the fact that as shown by Peña and Rodriguez (2002, Table 2) the small sample performance is acceptable in some cases, the more general use of tests based on the gamma approximation

Table 2
Asymptotic probability corresponding to the upper 5% quantile of the empirical distribution of $\hat{D}_m$ with $m = 50$ for the first-order autoregressive model with parameter ϕ and series length n

n	ϕ			
	−0.8	−0.4	0.4	0.8
100	0.423	0.433	0.439	0.445
200	0.179	0.161	0.170	0.181
500	0.093	0.092	0.081	0.079
1000	0.060	0.059	0.051	0.062

Each entry in the table is based on 10^4 simulations.

cannot be recommended. The reader may investigate the accuracy of the gamma approximation for a particular m and ARMA(p, q) using our `gvtest` package that is available on CRAN.

Another serious limitation to the use of D_m is that it is frequently undefined. This happens because $\ddot{R}_m$ is not always positive definite (McLeod and Jimenéz, 1984). When we tried to replicate Table 4 in Peña and Rodriguez (2002) we frequently found cases where $\ddot{R}_m$ was not positive definite. For example, for model 7 listed in this table this happens about 25% of the time with $m = 10$. This problem occurred with many other models listed in this table. Although only very short time series were used in Table 4, this problem also occurs with longer time series particularly when m is also larger. For this reason, it is better to concentrate on the original $\hat{D}_m$ statistic.

3. Tests based on $\hat{D}_m$

In a modern high-level computing environment, such as R or *Mathematica*, it is not difficult to evaluate the asymptotic distribution and so a test based directly on this asymptotic distribution might seem to be preferable. Unfortunately, the convergence of $\hat{D}_m$ to its asymptotic distribution is often very slow. In Table 2, we evaluated the asymptotic distribution corresponding to the upper 5% point of the empirical distribution of $\hat{D}_m$ for the first-order autoregressions with series lengths $n = 100,\ 200,\ 500$ and 1000. Not until n is very large is the asymptotic distribution reliable. Fig. 2 shows a QQ plot for these simulations which shows that the discrepancy between the empirical and asymptotic distribution increases at the larger quantiles. The asymptotic distribution tends to understate the actual finite-sample significance level while the gamma approximation errors in the opposite direction.

For these reasons, a Monte-Carlo test procedure (Gentle, 2002, Section 2.3) is recommended when $n < 1000$. This procedure is quite practical on typical computers now available.This test is essentially equivalent to a parametric bootstrap test (Davison and Hinkley, 1997, Ch.4). The steps in this procedure are indicated below:

1. After fitting model in Eq. (1) obtain $\hat{D}_m$.
2. Select the number of Monte-Carlo simulations, N. Typically $100 \leqslant N \leqslant 1000$.
3. Simulate the model in Eq. (1) using the estimated parameters obtained in Step 1 and obtain $\hat{D}_m$ after estimating the parameters in the simulated series.
4. Repeat Step 3 N times counting the number of times k that a value of $\hat{D}_m$ greater than or equal to that in Step 1 has been obtained.
5. The P-value for the test is $(k + 1)/(N + 1)$.
6. Reject the null hypothesis if the P-value is smaller than a predetermined significance level.

It should be noted that nuisance parameters are present in our proposed procedure and this could cause size distortion in the Monte-Carlo test (Jöckel, 1986). The empirical size of the Monte-Carlo test for the first-order autoregressive model was investigated by simulation. The results were summarized in Table 3 and it is seen that the empirical sizes are very close to their nominal level. In general, it has been shown (Dufour, 2006; Dufour and Khalaf, 2001) that if consistent estimators are used then the Monte-Carlo test produces an asymptotically correct size as $n \to \infty$. The asymptotic validity of the Monte-Carlo test follows immediately.

Alternatively since the gamma approximation is asymptotically correct as both n and m get large, it follows that $\hat{D}_m$ is asymptotically pivotal and hence does not depend on nuisance parameters.

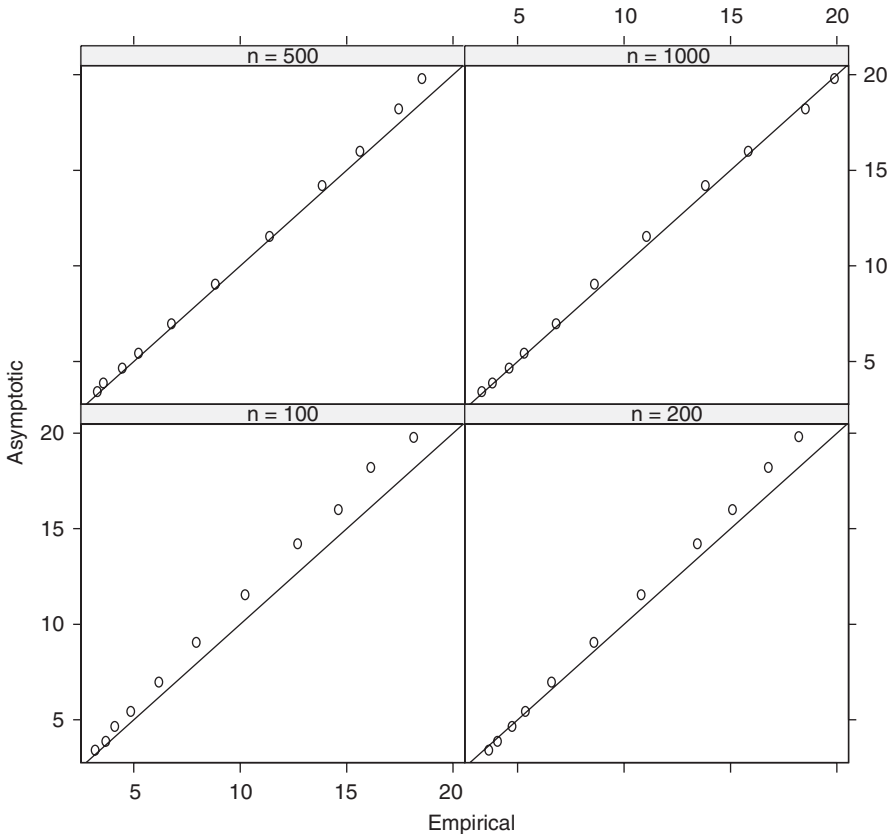

Fig. 2. Slow convergence of empirical distribution of $\hat{D}_m$. The AR(1) with parameter $\phi_1 = 0.5$ was simulated 10^3 times for series of lengths $n = 100, 200, 500, 1000$ and the empirical distribution of $\hat{D}_m$ with $m = 20$ was compared with its asymptotic distribution using a QQ plot with quantiles corresponding to 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.7, 0.9, 0.95, 0.98, 0.99.

Table 3
Empirical significance levels of $\hat{D}_m$ under first-order autoregressive models

	$\phi = 0.1$	$\phi = 0.3$	$\phi = 0.5$	$\phi = 0.7$	$\phi = 0.9$
$\alpha = 0.05$					
$m = 10$	0.050	0.049	0.059	0.055	0.049
$m = 20$	0.049	0.047	0.046	0.050	0.058
$\alpha = 0.01$					
$m = 10$	0.016	0.014	0.011	0.012	0.008
$m = 20$	0.009	0.012	0.014	0.014	0.012

The empirical power for the MC test is based on 1000 simulations. Each MC test also used 250 simulations.

4. Empirical power comparisons

In many cases the Monte-Carlo test outperforms the D_m test based on the gamma approximation. As an illustration, in Table 4 the simulation results for the Monte-Carlo test are compared with D_m test for the GARCH models Peña and Rodriguez (2002, Table 12). We see that the Monte-Carlo test has always has higher power than the gamma approximation in Table 4.

Peña and Rodriguez (2002) indicated that the advantage of their test over the Ljung–Box test, denoted as Q_m, may disappear in heteroscedastic data with long persistence. The results in Table 5 confirm this fact. For example, $\hat{D}_m$ has

Table 4
Power comparison of MC test and D_m test for GARCH time series

Model	n	$m=12$		$m=24$		$m=32$	
		$\hat{D}_m$	D_m	$\hat{D}_m$	D_m	$\hat{D}_m$	D_m
A	250	0.387	0.268	0.397	0.244	0.381	0.213
B	250	0.878	0.821	0.859	0.782	0.841	0.731
A	500	0.574	0.522	0.566	0.501	0.566	0.479
B	500	0.992	0.986	0.990	0.981	0.987	0.973
A	1000	0.843	0.807	0.843	0.802	0.829	0.776
B	1000	1.000	1.000	1.000	1.000	1.000	1.000

The empirical power for the MC test is based on 1000 simulations. Each MC test also used 1000 simulations. The empirical power reported by Peña and Rodriguez (2002, Table 9) is shown in the column D_m. Models A and B refer to the two GARCH models used by Peña and Rodriguez (2002).

Table 5
Empirical power comparison of Monte-Carlo test using $\hat{D}_m$ and Ljung–Box test, Q_m, for fractional noise time series with $n=256$, 512 and $d=0.2$, 0.3

d	n	$m=5$	$m=10$	$m=20$	$m=30$	$m=40$
0.2	256	0.283/0.274	0.274/0.247	0.227/0.202	0.192/0.176	0.171/0.153
0.3	256	0.539/0.517	0.540/0.486	0.464/0.411	0.419/0.329	0.374/0.307
0.2	512	0.620/0.614	0.609/0.557	0.547/0.455	0.476/0.406	0.446/0.377
0.3	512	0.888/0.880	0.894/0.848	0.851/0.804	0.823/0.736	0.778/0.708

The empirical power is based on 1000 simulations and each Monte-Carlo test also uses 1000 replications. The first entry corresponds to $\hat{D}_m$ and the second to Q_m.

Table 6
Empirical power of Monte-Carlo tests $\hat{D}_m/Q_m$ when a first-order autoregressive model is fit to various indicated autoregressive-moving average models denoted by models 1–12 in Peña and Rodriguez (2002, Table 3)

Model	$m=10$	$m=20$
1	0.464/0.259	0.361/0.175
2	0.988/0.705	0.974/0.539
3	0.993/0.757	0.990/0.574
4	0.584/0.414	0.472/0.303
5	0.584/0.414	0.472/0.303
6	0.798/0.490	0.702/0.361
7	1.000/0.981	1.000/0.846
8	1.000/0.821	0.998/0.621
9	0.246/0.169	0.196/0.130
10	0.876/0.743	0.820/0.575
11	0.252/0.119	0.192/0.083
12	0.989/0.664	0.979/0.472

The series length was $n=100$ and 1000 simulations were used for each test.

power 0.888 for $m=5$ and $d=0.3$ with respect to a power of 0.880 for Q_m. However, it is interesting to note that, as can be seen in Table 5, the difference in power increases as n and m increase, so that, for example, when $n=512$, $m=40$, $\hat{D}_m$ is about $100\times(0.446-0.377)/0.377\doteq 18\%$ more powerful than Q_m.

In another simulation experiment, shown in Table 6, we compared the power of Monte-Carlo tests using both $\hat{D}_m$ and Q_m for 12 models examined by Peña and Rodriguez (2002, Table 3). Overall our results are similar to those reported by Peña and Rodriguez (2002, Table 3). There are some differences though and this is due the limitations discussed in Section 2. Specifically, incorrect size when the gamma approximation was used or bias in the simulations caused by the fact that D_m frequently does not exist. We also investigated Monte-Carlo tests based on the Ljung–Box test, Q_m.

Table 7
Comparison of p-values for portmanteau tests

	m	20	30	40	50
Ninemile	Q_m	0.9%	8.0%	22.3%	32.2%
	$\hat{D}_m$	0.4%	0.5%	1.1%	1.4%
	m	5	10	15	20
Series E	Q_m	4.6%	10.0%	9.0%	21.7%
	$\hat{D}_m$	2.3%	4.3%	4.3%	5.3%

The Ljung–Box Q_m test and the Monte-Carlo $\hat{D}_m$ test are compared for the ARMA(2, 1) model fit to the Ninemile time series and the AR(2) model fit to Series E.

As shown in Table 6, the Monte-Carlo test using $\hat{D}_m$ outperforms this test. The empirical power for the Monte-Carlo test for the Box–Pierce test (Box and Pierce, 1970) was also computed but it was not significantly different from the results for the Q_m test.

5. Illustrative examples

Our Monte-Carlo test is implemented in an R package available on CRAN, `GVTest`. Hipel and McLeod (1978) fit an ARMA(2, 1) model a tree-ring time series denoted by `Ninemile`. There were $n = 771$ annual values. As can be seen from Table 7, if one uses the Ljung–Box portmanteau test with $m = 40$, the model appears adequate although for $m = 20$ the test does strongly suggest model inadequacy. In this case the $\hat{D}_m$ Monte-Carlo test provides a clearer indication since it indicates to reject for $m = 20, 30, 40, 50$.

Fitting an AR(2) model to the sunspot series in Box et al. (1994, Series E) we again found that the Ljung–Box test suggests the model is adequate but the Monte-Carlo $\hat{D}_m$ test indicates model inadequacy. Note that with $m = 5$, the Ljung–Box test does have a p-value of about 5% but usually m is taken to be larger since only for large enough m is the covariance matrix idempotent. For this reason, the result for $m = 5$ might be discounted. When a Monte-Carlo test is used, no such difficulties arise. The results are summarized in Table 7.

6. Concluding remarks

The implementation of the generalized variance portmanteau test statistic suggested by Peňa and Rodriguez (2002) is unsatisfactory because it frequently does not exist and there are important limitations to the gamma approximation. These difficulties are rectified by using a Monte-Carlo test. Further simulation experiments have indicated that the Monte-Carlo portmanteau test using simulated Gaussian innovations often works well even when the true distribution is non-normal. This is the case for thicker tail distributions such as double exponential and the t distribution on 5 df.

Acknowledgement

A.I. McLeod acknowledges with thanks a Discovery Grant Award from NSERC.

References

Box, G.E.P., Pierce, D.A., 1970. Distribution of the residual autocorrelation in autoregressive integrated moving average time series models. J. Amer. Statist. Assoc. 65, 1509–1526.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis: Forecasting and Control. third ed. Holden-Day, San Francisco.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge Unviersity Press, Cambridge.

Dufour, J.-M., 2006. Monte Carlo tests with nuisance parameters: a general approach to finite-sample inference and nonstandard asymptotics in econometrics. J. Econometrics, forthcoming.

Dufour, J.-M., Khalaf, L., 2001. Monte-Carlo test methods in econometrics. In: Companion to Theoretical Econometrics. Blackwell, Oxford, Chapter 23, 494–519.

Gentle, J.E., 2002. Elements of Computational Statistics. Springer, New York.

Hipel, K.W., McLeod, A.I., 1978. Preservation of the rescaled adjusted range, Part 2, Simulation studies using Box-Jenkins models. Water Resources Research 14, 509–516.

Imhof, J.P., 1961. Computing the distribution of quadratic forms in normal variables. Biometrika 48, 419–426.
Jöckel, K.F., 1986. Finite sample properties and asymptotic efficiency of Monte Carlo tests. Ann. Statist. 14, 336–347.
Li, W.K., 2004. Diagnostic Checks in Time Series. Chapman & Hall/CRC, New York.
Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika 65, 297–303.
McLeod, A.I., 1978. On the distribution of residual autocorrelations in Box-Jenkins models. J. Roy. Statist. Soc. B 40, 396–402.
McLeod, A.I., Jiménez, C., 1984. Nonnegative definiteness of the sample autocorrelation function. Amer. Statist. 38, 297–298.
Peňa, D., Rodriguez, J., 2002. A powerful portmanteau test of lack of fit for time series. J. Amer. Statist. Assoc. 97, 601–610.