Goodness-of-fit tests of normality for the innovations in ARMA models

(abbreviated title: Testing the residuals in ARMA)

Gilles R. Ducharme

and

Pierre Lafaye de Micheaux

Laboratoire de probabilités et statistique, cc051
Université Montpellier II
Place Eugène Bataillon
34095, Montpellier, Cedex 5
France

Technical report

Février 2002

Key Words: ARMA process, Gaussian white noise, Goodness-of-fit test, Normality of residuals, Smooth test.

Summary

In this paper, we propose a goodness-of-fit test of normality for the innovations of an ARMA(p, q) model with known mean or trend. The test is based on the data driven smooth test approach and is simple to perform. An extensive simulation study is conducted to study the behavior of the test for moderate sample sizes. It is found that our approach is generally more powerful than existing tests while holding its level throughout most of the parameter space and thus, can be recommended. This meshes with theoretical results showing the superiority of the data driven smooth test approach in related contexts.

1 Introduction

Let $(Y_t,\, t \in \mathbb{Z})$ be a stationary process. In this paper, we concentrate on the case where $E(Y_t)$ is known or has been estimated using information outside of the data set. Thus, without loss of generality, we set $E(Y_t) = 0$. Consider the framework where $(Y_t,\, t \in \mathbb{Z})$ obeys the causal and invertible finite order ARMA(p, q) model

$$Y_t - \boldsymbol{\varphi}^{\mathsf{T}} \boldsymbol{Y}_{t-1}^{(p)} = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\epsilon}_{t-1}^{(q)} + \epsilon_t \tag{1.1}$$

where $(\epsilon_t,\, t \in \mathbb{Z})$ is an innovation process of random variables with mean 0 and autocovariance $E(\epsilon_t \epsilon_{t+h}) = \sigma^2 < \infty$ (unknown) if $h = 0$ and 0 otherwise and where

$$\boldsymbol{\varphi} = \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_p \end{bmatrix},\ \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_q \end{bmatrix},\ \boldsymbol{Y}_{t-1}^{(p)} = \begin{bmatrix} Y_{t-1} \\ \vdots \\ Y_{t-p} \end{bmatrix},\ \boldsymbol{\epsilon}_{t-1}^{(q)} = \begin{bmatrix} \epsilon_{t-1} \\ \vdots \\ \epsilon_{t-q} \end{bmatrix}.$$

A sample $\{Y1, ..., Y_T\}$ is observed and model (1.1) is fitted by standard methods, e.g. the unconditional Gaussian maximum likelihood approach (see Brockwell & Davis (1991), p. 256-257), yielding the estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\varphi}}^{\mathsf{T}}, \hat{\boldsymbol{\theta}}^{\mathsf{T}}, \hat{\sigma})^{\mathsf{T}}$ of $\boldsymbol{\beta} = (\boldsymbol{\varphi}^{\mathsf{T}}, \boldsymbol{\theta}^{\mathsf{T}}, \sigma)^{\mathsf{T}}$.

If it can be safely assumed that the $(\epsilon_t,\, t \in \mathbb{Z})$ generating the Y_t's is of a given distributional type, in particular independent identically distributed (*i.i.d.*) normal (Gaussian) random variables, then better inference can be drawn from the fitted model. For example, such an assumption is helpful in obtaining accurate confidence or tolerance bounds for a predicted Y_{T+h}. Moreover, under this Gaussian assumption, $\hat{\boldsymbol{\beta}}$ is asymptotically efficient. Thus it is important to be able to test a null hypothesis of the form

$$H_0 : \text{ the } \epsilon_t\text{´s are } i.i.d. \sim N(0, \sigma^2). \tag{1.2}$$

As pointed out by Pierce & Gray (1985) and Brockett et al. (1988), other reasons may motivate a test of (1.2). One such reason is to check the validity of the structural part of (1.1). Indeed, the process of fitting a model to data often reduces to that of finding the model whose residuals behave most like a sample of *i.i.d.* Gaussian variables. In this context, rejection of (1.2) may indicate lack-of-fit of the entertained ARMA model. We will not elaborate further here on this possibility and assume, in the sequel, that model (1.1) is not underspecified. Note however that there exist specific tests for detecting lack-of-fit (for a recent review, see Koul & Stute (1999).

For the problem of testing (1.2), the few tests available fall roughly into two groups. Tests in the first group use the fact that for the ARMA (p, q) models considered here, normality of the Y_t's induces normality of the ϵ_t's and vice-versa. Thus a test of the hypothesis that a process (Y_t, $t \in \mathbb{Z}$) is Gaussian (Lomnicki (1961); Hinich (1982); Epps (1987)) can serve for problem (1.2). This presents the advantage of not requiring the values of p and q. But Gasser (1975) and Granger (1976) have shown, and Lutkepohl & Schneider (1989) have confirmed by simulation, that this approach may lose much power. This is because the central limit theorem forces the Y_t's to be close to normality even when (1.2) is false. Moreover, the adaptation of standard normality tests to dependent data is not an easy task. A small simulation study by Heuts & Rens (1986) has shown that, because of the serial correlation between the Y_t's, the finite null behavior of standard normality tests based on the empirical distribution function (EDF) of the Yt's is different from what is obtained under *i.i.d.* data. The same problem appears for tests based on the third and fourth moments of Y_t (see Lomnicki (1961); Lutkepohl & Schneider (1989)) or Pearson´s chi-square (Moore (1982)).

It thus appears better, when there are reasons to believe that a given ARMA(p, q) model holds, to "inverse filter" the data and compute the residuals $\hat{\epsilon}_t$ after fitting model (1.1). These can then be injected in some test for normality. Tests in the second group are based on this idea and some examples are listed in Hipel & McLeod (1994). However, these and other authors use such tests in conjunction with critical values for *i.i.d.* data. Since the residuals of an ARMA model are dependent, the null distribution of standard test statistics may be affected and critical values for *i.i.d.* data may no longer be valid. It turns out that for AR models, there is theoretical proof that this dependence disturbs only slightly the critical values, at least when T is large. For an AR(p) model with unknown $E(Y_t)$, Pierce & Gray (1985) has shown that the asymptotic null distribution of any test statistic based on the EDF of the residuals coincides with that of the same statistic for *i.i.d.* data with mean and variance unknown. Thus one can drop the residuals from an AR(p) model into any of the standard EDF-based tests (Kolmogorov-Smirnov, Anderson-Darling) and if T is large, use the critical values given, for example, in Chapter 4 of D'Agostino & Stephens (1986), to obtain an asymptotically valid test strategy. In the same vein, Lee & Na (2002) have recently adapted the Bickel-Rosenblatt test to this AR setting. Beiser (1985) has found that for the AR(1) model, tests based on the skewness or kurtosis coefficients of the residuals (D'Agostino & Stephens (1986), p. 408) in conjunction with the critical points derived for *i.i.d.* data produce valid levels if T is large and the AR-parameter is not too close to its boundary. This has been confirmed by Lutkepohl & Schneider (1989). See also Anděl (1997).

For the general ARMA model, much less information is available. Ojeda et al. (1997) show that tests based on quadratic forms in differences between sample moments and expected values of certain non-linear functions of the sample have the same asymptotic distribution under the ARMA model than under *i.i.d.* data. This suggests that a generalization of Pierce & Gray (1985) theorem to ARMA models could hold although, to our knowledge, no proof of this has been published. Otherwise, the practice recommended in many textbooks (see for example, Brockwell & Davis (1991), p. 314; Hipel & McLeod (1994), p. 241) is to use standard normality tests in conjunction with critical values for *i.i.d.* data.

In this paper, we develop some tests specifically designed for problem (1.2) in the ARMA(p, q) context. Our approach is based on the smooth test paradigm introduced by Neyman (1937) and improved by the data driven technology introduced by Ledwina (1994) to select the best order for the test. This has been shown in the *i.i.d.* case to offer many advantages, both theoretically and empirically, over other approaches. In particular, the test statistic we recommend for problem (1.2) is easy to compute with an asymptotic χ^2 distribution that can be corrected in finite samples to yield close to nominal levels. Moreover, as a byproduct of the procedure, diagnostic information is available that helps in understanding which aspects of the null hypothesis are not supported by the data.

Note that we concentrate here on the development of valid tests along this paradigm and do not dwell into their theoretical properties (i.e. local power and asymptotic efficiency). We also stress that the tests proposed here are valid solely for the case where $E(Y_t)$ is assumed known. The case where an unknown trend is present in (1.1) requires a special treatment and is the object of current research.

The paper is organized as follows. In Section 2, we develop the smooth goodness-of-fit test in the ARMA(p, q) context of (1.1). In Section 3, we describe the data-driven technology that allows to "fine tune" the test by choosing a good value for its order. In Section 4, a comprehensive Monte-Carlo study is conducted for some values of (p, q) to study the behavior of the proposed tests under the null hypothesis and compare their power to some competitors. It emerges that, under the null hypothesis, one of our data driven smooth tests holds its level over most of the parameter space and, under the alternatives studied, is in general more powerful than other methods. It can thus be recommended as a good tool for problem (1.2). An example concludes the paper.

2 Smooth test of normality in the ARMA context

Let $\Phi(.)$ denote the cumulative distribution function of the $N(0,1)$ distribution with density $\phi(.)$ and consider the random variables $U_t = 2\Phi(\epsilon_t/\sigma) - 1$ with density $g(.)$. Under H_0 of (1.2), the U_t's are *i.i.d.* U[-1, 1] random variables so that (1.2) reduces to testing whether $g(u) = 1/2$ on [-1, 1]. The ϵ_t's are unobserved and the test must be based on residuals. Since the process $(Y_t, t \in \mathbb{Z})$ is invertible, we have

$$\epsilon_t = -\sum_{r=0}^{\infty} \delta_r Y_{t-r} \tag{2.1}$$

where the δ_r's are functions of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ (see (A.2), (A.3) of Appendix A). Let $\hat{\delta}_r$ be the gaussian maximum likelihood estimator (*m.l.e.*) of δ_r under (1.2), obtained by plugging in the *m.l.e.* $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\varphi}}$ under H_0. We define the residuals of the fitted ARMA model by

$$\hat{\epsilon}_t = -\sum_{r=0}^{\infty} \hat{\delta}_r Y_{t-r}. \tag{2.2}$$

Of course, some scheme must be used in practice to compute these residuals, for example by taking $Y_t = 0$ if $t < 1$. Note that these residuals are not the only ones that can be defined for ARMA models (see Brockwell & Davis (1991), Section 9.4 for an alternative) but the definition above is convenient for the following derivation. Consider $\hat{U}_t = 2\Phi(\hat{\epsilon}_t/\hat{\sigma}) - 1$, $t = 1, ..., T$, which approximate the U_t's but have a complicated covariance structure. The problem is to construct, in spite of this, a valid goodness-of-fit test of (1.2).

To this end, let $\{L_k(.), k \geq 0\}$ be the normalized (over [-1, 1]) Legendre polynomials (Sansone (1959)) with $L_0(.) \equiv 1$ satisfying the orthonormality relation

$$\frac{1}{2}\int_{-1}^{1} L_k(x)L_j(x)dx = 1 \text{ if } k = j \text{ and } 0 \text{ otherwise.} \tag{2.3}$$

For some integer $K \geq 1$, consider the density defined on [-1, 1] by

$$g_K(u; \boldsymbol{\omega}) = c(\boldsymbol{\omega}) \exp\left\{\sum_{k=1}^{K} \omega_k L_k(u)\right\} \tag{2.4}$$

where $c(\boldsymbol{\omega})$ is a normalizing constant such that $c(\mathbf{0}) = 1/2$. In the classical smooth test paradigm, (2.4) is referred to as the K-th order alternative to the U[-1, 1] with $g_K(.;0)$ being the U[-1, 1] density. Thus, if $g(u)$ can be closely approximated by (2.4), problem (1.2) reduces to that of testing H_0: $\boldsymbol{\omega} = \mathbf{0}$. For this, we use the following route that leads essentially to Rao's score test. Let $\boldsymbol{L}_t = (L_1(U_t), ..., L_K(U_t))^\intercal$, $\hat{\boldsymbol{L}}_t = (L_1(\hat{U}_t), ..., L_K(\hat{U}_t))^\intercal$ and

$$\overline{\hat{\boldsymbol{L}}} = T^{-1} \sum_{t=1}^{T} \hat{\boldsymbol{L}}_t. \tag{2.5}$$

Under H_0, $\boldsymbol{L}_t$ has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}_K$, the K-th order identity matrix. Under alternative (2.4), these moments will differ and (2.5) can be used to capture departures from the U[-1, 1] in the "direction" of $g_K(.;\boldsymbol{\omega})$. This suggests as a test statistic a quadratic form in $\overline{\hat{\boldsymbol{L}}}$. In order to complete the test strategy, we need the null asymptotic distribution of (2.5). This is given in the following theorem.

Theoreme 2.1. *Consider the causal and invertible ARMA(p, q) process of (1.1) where we assume in addition that the polynomials $1 - \varphi_1 z - ... - \varphi_p z^p$ and $1 + \theta_1 z + ... + \theta_q z^q$ have no common zeroes. Under H_0, we have*

$$\sqrt{T}\overline{\hat{\boldsymbol{L}}} \xrightarrow{L} N_K\left(\mathbf{0}, \boldsymbol{I}_K - \frac{1}{2}\boldsymbol{b}_K\boldsymbol{b}_K^\intercal\right) \tag{2.6}$$

where $\boldsymbol{b}_K = (b_1, ..., b_K)^\intercal$, with $b_k = \int_{\mathbb{R}} L_k(2\Phi(x) - 1)x^2\phi(x)dx$. Hence, under H_0, the smooth test statistic

$$\mathcal{R}_K = T\overline{\hat{\boldsymbol{L}}}^\intercal \left(\boldsymbol{I}_K - \frac{1}{2}\boldsymbol{b}_K\boldsymbol{b}_K^\intercal\right)^{-1} \overline{\hat{\boldsymbol{L}}} \xrightarrow{L} \chi^2_K.$$

Proof. We present here the general outline of the argument. All details are confined to the appendices. Let

$$\mathcal{I}_{\boldsymbol{\beta}} = Var\left[\frac{\partial}{\partial\boldsymbol{\beta}} Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right)\right]$$

be Fisher's information matrix for $\boldsymbol{\beta}$. From standard results (see Gouriéroux & Monfort (1995), p.325), we have,

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathcal{I}_{\boldsymbol{\beta}}^{-1} \frac{\partial}{\partial\boldsymbol{\beta}} \left[Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right)\right] + o_P(1).$$

Since $\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=O_P(T^{-1/2})$, a Taylor expansion yields

$$\sqrt{T}\overline{\hat{\boldsymbol{L}}}=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{L}_t+\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{L}_t\right]\sqrt{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+o_P(1). \tag{2.7}$$

The first term on the right hand side of (2.7) converges to a $N_K(\mathbf{0},\boldsymbol{I}_K)$. Moreover, it is shown in Appendix A that

$$\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{L}_t\right]\xrightarrow{P}\left[\underset{K\times(p+q)}{\mathbf{0}},-\frac{1}{\sigma}\boldsymbol{b}_K\right]=-\mathcal{J}_K. \tag{2.8}$$

Hence,

$$\sqrt{T}\overline{\hat{\boldsymbol{L}}}=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{L}_t-\frac{1}{\sqrt{T}}\mathcal{J}_K I_{\boldsymbol{\beta}}^{-1}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\beta}}\left[Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right)\right]+o_P(1)$$

$$=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{B}\boldsymbol{V}_t+o_P(1)$$

where $\boldsymbol{B}=(\boldsymbol{I}_K,-\mathcal{J}_K\mathcal{I}_{\boldsymbol{\beta}}^{-1})$ and

$$\boldsymbol{V}_t=\left(\boldsymbol{L}_t^{\mathsf{T}},\frac{\partial}{\partial\boldsymbol{\beta}^{\mathsf{T}}}\left[Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right)\right]\right)^{\mathsf{T}}.$$

Now, it is shown in Appendix B that, under H_0, $E(\boldsymbol{V}_t)=\mathbf{0}$ and $Var(\boldsymbol{B}\boldsymbol{V}_t)=\boldsymbol{I}_K-\boldsymbol{b}_K\boldsymbol{b}_K^{\mathsf{T}}/2$. The central limit theorem yields (2.6). □

It is possible to write $\mathcal{R}_K$ in a form that makes it easier to use. A Cholesky decomposition of $\left(\boldsymbol{I}_K-\boldsymbol{b}_K\boldsymbol{b}_K^{\mathsf{T}}/2\right)$ yields $\left(\boldsymbol{I}_K-\boldsymbol{b}_K\boldsymbol{b}_K^{\mathsf{T}}/2\right)^{-1}=\boldsymbol{P}\boldsymbol{P}^{\mathsf{T}}$ with $\boldsymbol{P}=(p_{ij})$, an upper triangular matrix. Some algebra gives

$$p_{ij}=\begin{cases}0 & \text{if } i>j\\ \sqrt{\frac{2-\sum_{k=1}^{i-1}b_k^2}{2-\sum_{k=1}^{i}b_k^2}} & \text{if } i=j\\ \frac{b_ib_j}{\sqrt{\left(2-\sum_{k=1}^{j-1}b_k^2\right)\left(2-\sum_{k=1}^{j}b_k^2\right)}} & \text{if } j>i\end{cases}.$$

Thus

$$\mathcal{R}_K=\sum_{k=1}^{K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}L_k^*(\hat{U}_t)\right)^2$$

where

$$L_k^*(\hat{U}_t) = \sum_{l=1}^{k} p_{lk} L_l(\hat{U}_t). \tag{2.9}$$

Numerical integration gives $(b_2, b_4, ..., b_{10}) = (1.23281, 0.521125, 0.304514, 0.205589, 0.150771)$ with $b_k = 0$ if k is odd. This yields the first ten "modified" Legendre polynomials

$$\begin{aligned}
&L_1^*(u) = 1.73u,\\
&L_2^*(u) = 6.85u^2 - 2.28,\\
&L_3^*(u) = 6.61u^3 - 3.97u,\\
&L_4^*(u) = 19.91u^4 - 10.26u^2 - 0.56,\\
&L_5^*(u) = 26.12u^5 - 29.02u^3 + 6.22u,\\
&L_6^*(u) = 69.84u^6 - 81.84u^4 + 28.36u^2 - 3.06,\\
&L_7^*(u) = 103.84u^7 - 167.75u^5 + 76.25u^3 - 8.47u,\\
&L_8^*(u) = 260.07u^8 - 450.18u^6 + 247.18u^4 - 38.73u^2 - 1.11,\\
&L_9^*(u) = 413.92u^9 - 876.55u^7 + 613.58u^5 - 157.33u^3 + 10.73u,\\
&L_{10}^*(u) = 994.51u^{10} - 2250.43u^8 + 1782.83u^6 - 569.92u^4 + 67.54u^2 - 3.58.
\end{aligned}$$

Remark 2.1. *The test statistic derived here differs from the standard smooth test statistic for normality in the i.i.d. case given, for example, in Thomas & Pierce (1979). However, it turns out that if one computes, using their approach, the smooth test statistic for i.i.d. observations when the mean is known (case 2 in the terminology of D'Agostino & Stephens (1986), which corresponds, in the present context, to an ARMA(0, 0)), the resulting statistic coincides with our $\mathcal{R}_K$. Thus, we can slightly extend the finding of Pierce & Gray (1985) and state that neither the estimation of $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ nor the dependence of the Y_t's has any asymptotic impact on a smooth test of (1.2) in the present ARMA context. Of course, in pre-asymptotic situations these elements and the complexity of the model will affect the null distribution of $\mathcal{R}_K$. This will be further explored in the simulation study of Section 4.*

Remark 2.2. *Each term $T^{-1}\left(\sum_{t=1}^{T} L_k^*(\hat{U}_t)\right)^2$ is a component of the test statistic and has an asymptotic χ_1^2 distribution under H_0. When the null hypothesis is rejected, some of these components will be large. The simple structure of the first few polynomials in (2.9) allows some interpretation and helps in understanding what aspects of the normal distribution are not supported by the data. For example, the first component detects departure from* *symmetry under H_0 in the "direction" of asymmetry. This diagnostic analysis must be undertaken with some care however; see Henze (1997) for details.*

Remark 2.3. *The above methodology can in principle be applied to test other distribution than the normal. For location-scale densities, one needs to replace the normal distribution in the definition of U_t and follow the above derivation using the new null density. In the end, the structure of $\mathcal{R}_K$ will be similar to what is obtained above but the modified Legendre polynomials will change. For distributions with a sahpe parameter, the statistic will be more complex since the coefficients of these modified polynomials will in general depend on this unknown shape parameter that must be estimated.*

3 Choosing the order K of the alternative

Before applying the test strategy of Section 2, one must choose a value for K. Ideally, this choice should be made so that members of the embedding family $g_K(.;\boldsymbol{\omega})$ of (2.4) provide a good approximation to any plausible density $g(.)$ of U_t under the alternative. If K is taken too small, this approximation may be crude and the test loses power. If K is taken too large, power dilution can occur since $g_K(.;\boldsymbol{\omega})$ encompasses unnecessary "directions".

In most practical cases, the user has only, at best, a qualitative idea of the plausible alternatives and no specific value of K emerges naturally. In the *i.i.d.* case, a first school of thoughts argues that, as a rule of thumb, one can use a trade-off value of K between 2 and 4.

Recently, Ledwina (1994) and Kallenberg & Ledwina (1997a,b) have proposed and explored for *i.i.d.* data a method to choose adaptively a value for K. At the first step, Schwarz (1978)´s criterion is used to zero in on the value $\hat{K}$ that seems best in view of the data at hand. The smooth test strategy is then applied using the statistic $\mathcal{R}_{\hat{K}}$. Extensive simulations have shown that, even for small sample sizes, this so-called "data-driven smooth test" can yield power close to what would be obtained if one knew the true form of the alternative and had chosen the best value of K accordingly.

Presently, this approach has been investigated for the *i.i.d.* case only. We now present an extension to the ARMA context. Choose two integers $1 \leq d \leq D$ and consider the set of statistics $(\mathcal{R}_d, ..., \mathcal{R}_D)$. We seek a rule that will select a good $\mathcal{R}_K$ in this set. Write

$$\hat{K} = \min\left[\underset{d\leq s\leq D}{Argmax}\{\mathcal{R}_s - sLog(T)\}\right] \tag{3.1}$$

and denote $\mathcal{R}_{\hat{K}}(d)$, the test statistic $\mathcal{R}_{\hat{K}}$ selected by (3.1) in $(\mathcal{R}_d, ..., \mathcal{R}_D)$.

Theoreme 3.1. *Under H_0, $\hat{K} \longrightarrow d$ in probability and thus, $\mathcal{R}_{\hat{K}}(d)$ is asymptotically χ^2_d .*

Proof. Set $e_k = (k-d)LogT$. For $k \geq d$, $P(\hat{K} = k) \leq P(\mathcal{R}_k > e_k)$. Now, since each $\mathcal{R}_k$ is asymptotically χ^2_k under H_0, as T increases

$$P(\mathcal{R}_k > e_k) \longrightarrow 0$$

when k > d. It follows that $P(\hat{K} = d) = 1 - P(\hat{K} \geq d+1) \longrightarrow 1$. $\square$

For finite sample sizes, the asymptotic null distribution of Theorem 3.1 may not provide a good approximation to that of $\mathcal{R}_{\hat{K}}(d)$ since there is a positive probability that $\hat{K} \geq d+1$. A simple correction has been developed by Kallenberg & Ledwina (1997a) when $d = 1$ (*i.i.d.* data). Because of the asymptotic independence between the components of $\mathcal{R}_k$, this correction can easily be extended to $d > 1$ and to the present ARMA context. Indeed,

$$P(\mathcal{R}_{\hat{K}}(d) \leq x) = P(\mathcal{R}_d \leq x, \hat{K} = d) + P(\mathcal{R}_{d+1} \leq x, \hat{K} = d+1) + P(\mathcal{R}_{\hat{K}}(d) \leq x, \hat{K} > d+1).$$

Since under H_0, $P(\hat{K} > d+1)$ is small if T is large, we are brought to neglect the last term and, for the remaining terms, to consider the case $D = d+1$. Then, $\hat{K} = d$ when

$$\left(T^{-1/2} \sum_{t=1}^{T} L^*_{d+1}(\hat{U}_t) \right)^2 \leq Log(T). \tag{3.2}$$

This left-hand side of (3.2) is asymptotically independent of $\mathcal{R}_d$ so we get

$$P(\mathcal{R}_d \leq x, \hat{K} = d) \approx P(\chi^2_d \leq x) P(\chi^2_1 \leq Log(T)).$$

Moreover $\hat{K} = d+1$ when inequality in (3.2) is reversed. Thus

$$P(\mathcal{R}_{d+1} \leq x, \hat{K} = d+1) \approx \int_{Log(T)}^{x} P(\chi^2_d < x - z) \frac{1}{\sqrt{2\pi z}} e^{-z/2} dz.$$

This leads to the following approximation, which can be solved for x by numerical integration

$$P(\mathcal{R}_{\hat{K}}(d) \leq x) \approx P(\chi^2_d \leq x) P(\chi^2_1 \leq Log(T)) + \int_{Log(T)}^{x} P(\chi^2_d < x - z) \frac{1}{\sqrt{2\pi z}} e^{-z/2} dz. \tag{3.3}$$

Some quantiles corrected through (3.3) are listed in Table 3.1.

	T	a = 0.10	a = 0.05	a = 0.01
d = 1	50	3.692	5.410	8.805
	100	3.275	5.201	8.703
	200	3.057	4.751	8.590
d = 2	50	5.466	7.137	10.807
	100	5.262	6.972	10.684
	200	5.043	6.796	10.558

Table 3.1: Some quantiles obtained from approximation (3.3)

One may get the feeling that this data driven approach replaces the problem of selecting K with that of selecting d and D. To answer this, Kallenberg & Ledwina (1997a,b) have studied a version of the above procedure where D is allowed to increase with T. In the *i.i.d.* case, they obtain rates connecting these quantities. These rates are theoretically interesting but do not help in practice in selecting a value for D. To obtain more insight, they have conducted extensive simulations. It turns out that the power levels off rapidly as D increases and there is little to be gained by choosing D above a certain threshold in the area of 10. As for the choice of d, again Kallenberg & Ledwina (1997a) briefly discuss this problem where it emerges that in their context $d = 1$ or 2 appears reasonable. In the simulation study of the next section we use both these values of d and take $D = 10$.

In closing this section, note that, by plotting $g_{\hat{K}}(.;\hat{\boldsymbol{\omega}})$ where $\hat{\boldsymbol{\omega}}$ is an estimate of $\boldsymbol{\omega}$, one can get an idea of the true shape of the density when the null hypothesis has been rejected. This can be helpful in finding a more appropriate distribution for the innovations.

4 Simulation Results

To get an idea of the behavior of our test statistics as compared to some competitors, a simulation study was conducted.

The general framework of the study is as follows. Samples $\{Y_t, t = 1, ..., T\}$ from various ARMA(p, q) models were generated with the innovations arising, in the first part of the simulation, from the normal distribution and, in the second part, from various alternatives. For each of these samples, we estimated the parameters of the model and computed some test statistics. This allowed to obtain approximations to their actual level and power. All programs are written in Fortran 77 and all subroutines listed below are from the Numerical Algorithms Group (NAG) MARK 16 Fortran library.

4.1 Levels

The first part of the simulation study was designed to see if the critical values obtained from the asymptotic χ^2 or by (3.3) can be relied upon in finite samples. For this, we have taken $T = 50$, 100 and 200 and restricted attention to the models ARMA(0, 2) (= MA(2)), ARMA(2, 0) (=AR(2)), ARMA(1, 2), ARMA(2, 1) and ARMA(2, 2). To generate ARMA(p, q) samples with Gaussian innovations, we used subroutine G05EGF to obtain initial values corresponding to a stationary reference vector. Subroutine G05EWF was then used to generate the T successive terms of the sample. These samples were submitted to subroutine G13DCF that returns estimates of the parameters of the model as well as residuals. The definition of these residuals, which is given at equation (9.4.1) in Brockett et al. (1988), differs from (2.2) but their numerical values are almost identical. These residuals were then injected in the various test statistics. The actual levels of the tests were computed for nominal levels $\alpha = 0.10$ and 0.05. In all our simulations, we used subroutine G05CBF to set the seeds used by the random number generators to a repeatable initial value. The values of these seeds are available upon request.

Regarding the values of the parameter $\boldsymbol{\beta}$, notice first that our test statistics are theoretically invariant to the choice of σ and any convenient value can be taken: we have chosen $\sigma = 1$. Numerically, this invariance holds only approximately because of the stopping rule in the maximization algorithm.

On the other hand, the finite distribution of our test statistics depends on the values of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$. To explore this effect, we have proceeded as follows. First, causality requires that, if $p = 1$, $\varphi_1 \in]-1, 1[$ while if $p = 2$, $\boldsymbol{\varphi}$ must lay in the triangular region $\triangle_{\boldsymbol{\varphi}} = \{(\varphi_1, \varphi_2)|\varphi_1 + \varphi_2 < 1, \varphi_2 - \varphi_1 < 1, |\varphi_2| < 1\}$ (Brockett et al. (1988), p. 110, ex.3.2). Similarly, invertibility implies that if $q = 1$, then $\theta_1 \in]-1, 1[$ while if $q = 2$, $\boldsymbol{\theta}$ must lay in $\bigtriangledown_{\boldsymbol{\theta}} = \{(-\theta_1, -\theta_2)|\theta_1 + \theta_2 < 1, \theta_2 - \theta_1 < 1 \text{ and } |\theta_2| < 1\}$. In addition, the polynomials $1 - \varphi_1 z$ when $p = 1$ and $1 - \varphi_1 z - \varphi_2 z^2$ when $p = 2$ must not have any common zeroes with $1+\theta_1 z$ when q = 1 and $1+\theta_1 z+\theta_2 z^2$ when q = 2.

For the AR(2) model, we have taken the values of $\boldsymbol{\varphi}$ in the following grid of 64 points $\{(-2.0 + 0.25j, -0.9 + 0.25k) \in \triangle_{\boldsymbol{\varphi}}|j, k \geq 0\}$. A similar grid was used for the MA(2) model. This permits an exploration of the stability of the tests, with respect to maintaining the proper critical level, over a good part of the parameter space. Note that some points of these grids are close to the boundary of stationarity but none is extremely close.

For the ARMA(1, 2), the grid over $\bigtriangledown_{\boldsymbol{\theta}}$ was reduced to the points $\{(-2.0 + 0.40j, -0.9 + 0.40k) \in \bigtriangledown_{\boldsymbol{\theta}}|j, k \geq 0\}$ while $\varphi_1 = -0.9 + 0.2j, j = 0, ..., 9$. This gives a set of 250 points on the parameter space of $(\varphi_1, \boldsymbol{\theta})$. For the ARMA(2, 1) model, the same was done but with $\boldsymbol{\varphi}$ and θ_1 instead. Finally, for the ARMA (2, 2) model, points $(\boldsymbol{\varphi}, \boldsymbol{\theta})$ satisfying the "no common zeroes" condition stated above were taken in $\{(-1.95 + 0.45j, -0.85 + 0.45k) \in \triangle_{\boldsymbol{\varphi}}|j, k \geq 0\} \cup \{-(-1.95 + 0.45j, -0.95 + 0.45k) \in \bigtriangledown_{\boldsymbol{\theta}}|j, k \geq 0\}$. This yields 294 $(\boldsymbol{\varphi}, \boldsymbol{\theta})$ parameter points.

For each of these parameter points, 10000 samples of size T were generated as described above. The estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ can be a difficult non-linear problem. Subroutine G13DCF has an output parameter (IFAIL) that indicates whether a problem in obtaining the *m.l.e.* has been encountered. In our simulations, the most common problem flagged by this parameter was IFAIL = 5 and to a far lesser degree IFAIL = 6 and IFAIL = 8. IFAIL = 5 means that the conditions for a solution of the maximization process had not all been met but a point at which the log-likelihood takes a larger value could not be found. This indicates that the algorithm has probably found the solution, as far as the accuracy of the machine permits. IFAIL = 6 and IFAIL = 8 appear when the solution is so close to the boundary of stationarity and invertibility that there are some problems in evaluating the Hessian matrix. In this case, the attained solution is as close as computationally feasible to the *m.l.e.*. In view of the fact that such problems arise from the finite accuracy of the computer, we decided to keep all generated samples in the simulation. However, we checked that applying our test statistics either to the samples for which no computational problem occurred or to all samples produced virtually identical results. Moreover, close inspection confirmed that the samples where a flag appeared lead to estimates of parameters and residuals not discernably different from those obtained when no problem were reported.

To summarize the results of this part of the simulation, the following approach was adopted. A 95% confidence interval for the true level of our tests when the number of replication is 10000 and $\alpha = 0.10$ (resp. 0.05), has length approximately 0.012 (resp. 0.008). This suggests that p-values in the interval (0.094, 0.106) do not differ markedly from the nominal 10%. Similarly, for $\alpha = 0.05$, 95% of the p-values are expected in the interval (0.046, 0.054). Thus the range of possible p-values was divided in 5 subintervals. For $\alpha = 10\%$, these are $I_1 = (0, 0.085)$, $I_2 = [0.085, 0.094)$, $I_3 = [0.094, 0.106)$, $I_4 = [0.106, 0.115)$ and $I_5 = [0.115, 1]$. For $\alpha = 0.05$, $I_1 = (0, 0.035)$, $I_2 = [0.035, 0.046)$, $I_3 = [0.046, 0.054)$, $I_4 = [0.054, 0.065)$ and $I_5 = [0.065, 1]$. For each model, the percentage of p-values falling into each interval was recorded. Table 4.1 reports the results for statistics $\mathcal{R}_3$ and $\mathcal{R}_{\hat{K}}(2)$ which, as discussed in Section 3, are representative of the two schools of thought for the choice of K. Note that the sum of each line may differ from 100% due to rounding errors.

Insert Table 4.1.a) about here

Insert Table 4.1.b) about here

From Table 4.1a), we can observe for $\mathcal{R}_3$ that the actual levels are concentrated on I_1, I_2 and I_3 and the levels of this test rarely overestimate the intended level. The mode of the distribution is located on I_2 for $T = 50$ and is shifted on I_3 as T increases. This lead, at worst, to slightly conservative tests. To appreciate this, the last column of Table 4.1 gives the smallest p-value recorded over the parameter points. For $\mathcal{R}_{\hat{K}}(2)$ (Table 4.1.b), the bulk of the distribution is concentrated on I_2, I_3 and I_4 with, in all cases, a mode centered on I_3. For this statistic, the minimal p-values are also closer to the nominal level (no maximal p-value was very far from the upper bound of I_4). It thus appears that correction (3.3) works adequately, at least for the models and sample sizes considered here. Not surprisingly, for both statistics the dispersion about I_3 decreases as T increases. Overall, the statistic that yields the best results in this part of the simulation is $\mathcal{R}_{\hat{K}}(2)$.

We also investigated what areas of the parameter spaces gave p-values in I_1. This is important in practice since, if a sample yields estimates in such an area, the test of (1.2) might be conservative and some care must be taken in interpreting the results. Intuitively, one can expect these points to be near the boundary of the parameter space. However, the pattern that emerges, which is very similar for both $\mathcal{R}_3$, and $\mathcal{R}_{\hat{K}}(2)$, is more precise. For AR(2) models, these points correspond mainly to positive (φ_1, φ_2) close to the right boundary of $\triangle_{\boldsymbol{\varphi}}$ and, to a lesser degree, to those with positive φ_1 and negative φ_2 but again close to that boundary. For MA(2) models, the situation is reversed, which is not surprising since $\bigtriangledown_{\boldsymbol{\theta}} = -\triangle_{\boldsymbol{\varphi}}$. For ARMA(2, 1), the points giving small p-values correspond to positive (φ_1, φ_2) combined with values of θ_1 close to -1. Again, for ARMA(1, 2) the situation is reversed and small p-values are associated with negative values of (θ_1, θ_2) with a value of φ_1 close to 1. Finally, for the ARMA(2, 2), the points that yield p-values in I_1 are mainly those with positive (φ_1, φ_2) and negative (θ_1, θ_2). As T increases, these points are associated with p-values that are still smaller than nominal, but tend to fall in I_2 and eventually in I_3.

We have also investigated the null behavior of some other tests that have been recommended in the time series literature to check assumption (1.2). We first considered the Anderson-Darling ($\mathcal{AD}$) test Pierce & Gray (1985) for case 2 (known mean) used in conjunction with the quantiles given for this case in D'Agostino & Stephens (1986) p. 122. A table similar to 4.1 was obtained, but we report here only the part about the ARMA(1,2) model (see Table 4.2). Our simulations show that, for large T this yields valid critical levels under the null hypothesis. This supports the conjecture that Pierce's theorem could be extended to ARMA models. We also studied a variant of the Shapiro-Wilk test known as the Weisberg & Bingham (1975) ($\mathcal{W}$B) test. This variant was chosen because the coefficients in the test statistic are easily obtained for any sample size by Bloom's approximation (see D'Agostino & Stephens (1986) p. 400, eq. 9.67) and leads, in the *i.i.d.* case, to powers similar to that of the Shapiro-Wilk test. In order to adapt this test statistic to our time series context with known mean, the denominator of equation (9.68) of D'Agostino & Stephens (1986) was replaced by $T\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the estimate of σ^2 returned by subroutine G13DCF. Up to the numerical accuracy of procedure G13DCF, this corresponds to the sum of square of the residuals (see Brockwell & Davis (1991), p.257 eq. 8.7.5). The quantiles for this test (or any variant of the Shapiro-Wilk test) are unknown in the present context. It turns out that our simulations show that they can be approximated by Monte Carlo using *i.i.d.* data, although we found no theoretical results supporting this. Thus, we have simulated 100000 samples from an ARMA(0,0) model and computed the empirical quantiles. For $T = 50$, 100 and 200, we got, for $\alpha = 10\%$, 0.920, 0.958 and 0.978 while for 5%, we found 0.899, 0.947 and 0.973. A third approach, the Jarque & Bera (1987) eq. (5) ($\mathcal{J}$B) test was also investigated. Although developed in the linear regression context, this test has been sometimes recommended in the time series literature (see Cromwell et al. (1994); Frances (1998)). A summary of the results for these tests in the ARMA(1, 2) model is given in Table 4.2. Also appearing in this table are the levels of the test based on $\mathcal{R}_{\hat{K}}(1)$.

Overall, the best tests as regards levels are $\mathcal{R}_{\hat{K}}(2)$ followed by $\mathcal{R}_{\hat{K}}(1)$ and then $\mathcal{R}_3$, $\mathcal{AD}$ and $\mathcal{WB}$. In general, the AD test yields distributions of p-values in between those of $\mathcal{R}_3$ and $R_{\hat{K}}(1)$. More troublesome is the fact that this test may underestimate much more the intended level, as can be seen by the minimal p-values (last column of Table 4.2) that were encountered on the grid of parameter points. This indicates that the levels of the $\mathcal{AD}$ test are much less stable. The $\mathcal{WB}$ test exhibits a similar behavior. On the other hand, there appears to be a problem with the $\mathcal{JB}$ test as the 10% quantile, obtained from the asymptotic χ^2_2 approximation, is vastly in error and produces levels that are about half of the intended. The 5% quantile yields better results but they still underestimate the intended level at $T = 200$. Further simulations with this test show that levels improve slightly when $T = 500$, indicating that the convergence to the asymptotic χ^2_2 is very slow. The $\mathcal{JB}$ statistic is a version of the Bowman and Shenton test statistic (see Lutkepohl & Schneider (1989)) based on the empirical skewness and kurtosis of the residuals. For *i.i.d.* data, this statistic has a notoriously slow convergence toward its asymptotic distribution and for finite samples, corrected quantiles in the manner of Doornik & Hansen (1994) are required. The simulation results in Lutkepohl & Schneider (1989) tend to show that this is also the case for AR(1) and AR(2) models. Since the derivation of corrected quantiles in the present ARMA context is beyond the scope of the paper, we choose not to pursue further the investigation of the $\mathcal{JB}$ test.

Insert Table 4.2 about here

4.2 Power

The second part of our simulation was conducted to study the power of our tests and allow comparison with the competitors mentioned above. For this, we restricted attention to the case where the innovations are *i.i.d.* and generated samples $\{Y_t, t = 1, ..., T\}$ according to model (1.1) from various alternatives to the normal distribution. These alternatives were taken as the densities listed in Table V of Kallenberg & Ledwina (1997b) as defined on p.113 and centered so that their expectation is zero. They comprise a large range of departure from the normal distribution both in skewness, kurtosis and shape.

Generating ARMA(p, q) samples $\{Y_t, t = 1, ..., T\}$ according to model (1.1) with non-Gaussian innovations needs some care and, for this task, we used the random shock method described by Burn (1987). More precisely, we used his algorithms IA 1 with $m = 50$ and SA 1 with an induction period of $M = 200$. Again, the parameters of the models were estimated by subroutine G13DCF. To allow a proper comparison of the various tests, we used for each model a set of parameters for which the p-values computed in the first part of the simulation were in I_3 for all tests. More precisely we took: ARMA(2, 1): $(\boldsymbol{\varphi}, \theta_1) = (-0.8, -0.1, 0.7)$, ARMA(1, 2): $(\varphi_1, \boldsymbol{\theta}) = (-0.7, 0.4, 0.5)$ and ARMA(2, 2): $(\boldsymbol{\varphi}, \boldsymbol{\theta}) = (-1.05, -0.4, 0.15, 0.85)$. Also we took $T = 50$ and 100. For each combination of sample size, model and alternative distribution, we generated 10000 samples and performed the various tests. From there, empirical powers were computed.

Table 4.3 presents these empirical powers, with $\alpha = 10\%$, for the tests $\mathcal{R}_3$, $\mathcal{R}_{\hat{K}}(2)$, $\mathcal{AD}$ and $\mathcal{WB}$. Similar results were obtained for $\alpha = 5\%$. The tests $\mathcal{R}_3$ and $\mathcal{R}_{\hat{K}}(2)$ have a similar behavior with, overall, $\mathcal{R}_{\hat{K}}(2)$ being slightly better. Both these tests generally dominate the others. The $\mathcal{AD}$ approach often yields a power that is much lower than these two tests whereas $\mathcal{WB}$ generally lies somewhere in between. For *i.i.d.* data, the $\mathcal{WB}$ test, as a variant of the Shapiro-Wilk test, is considered among the best omnibus tests of normality. In ARMA situations, this does not seem to hold at the same degree. As an explanation, one can note that power decreases with the complexity of the model and that all powers reported in Table 4.3 are much lower than what is obtained from *i.i.d.* data (e.g. the ARMA(0,0) model). For both $\mathcal{R}_3$ and $\mathcal{R}_{\hat{K}}(2)$, the average loss in percentage points of power (*p.p.p.*) is 6.6 for the ARMA(2,1), 9.6 for the ARMA(1,2) and 13.2 for the ARMA(2,2). With such differences, it is perhaps not surprising that the behavior observed in the *i.i.d.* case differs from what has been observed in the present simulation.

We have also computed the power of the test based on $\mathcal{R}_{\hat{K}}(1)$. The tabulated results are not presented here for brevity. We found that, for $T = 50$ and symmetric alternatives, the test based on $\mathcal{R}_{\hat{K}}(1)$ yields slightly better power that $\mathcal{R}_{\hat{K}}(2)$ with an average increase of about 2.6 *p.p.p.*. For asymmetric alternatives, the situation is reversed and using $\mathcal{R}_{\hat{K}}(2)$ increases the power by 4.1 *p.p.p.* on the average. For $T = 100$, this small advantage vanishes, $\mathcal{R}_{\hat{K}}(2)$ being more powerful almost everywhere, with an average gain in power of about 1 *p.p.p.* for symmetric alternatives and 5.6 for asymmetric ones. This behavior of $\mathcal{R}_{\hat{K}}(1)$ is explained by the fact that for these alternatives, as with many others, $\mathcal{R}_1$ yields little, sometimes trivial, power. On the other hand, power as a function of K usually levels off at $\mathcal{R}_3$, and not infrequently at $\mathcal{R}_2$. This empirical observation is behind the rule of thumb stated in Section 3. Thus to have good power, the selection rule with $d = 1$ must give $\hat{K} \geq 3$, which may be difficult. By comparison, starting at $d = 2$ gives a better chance that $\hat{K} \geq 3$ when necessary without introducing power dilution when the best value of K is 2.

In view of the results of these simulations, we can recommend the use of $\mathcal{R}_{\hat{K}}(2)$ for testing (1.2) when $E(Y_t) = 0$. The levels are stable over most of the parameter points and close to nominal for moderate samples. Moreover, the obtained power is generally better than that of other tests that have been recommended in the time series literature. Finally, the test is very easy to apply: a FORTRAN program is available from the authors.

Insert Table 4.3.a) about here

Insert Table 4.3.b) about here

Insert Table 4.3.c) about here

5 An example

In the course of a study to forecast the amount of daily gas required, Shea (1987) has studied a bivariate time series of $T = 366$ points. The first component of this time series pertains to differences in daily temperature between successive days ($\bigtriangledown\tau_t$) and he found, after an iteration process of fitting and diagnostic checking, that the following MA(4) model could be entertained:

$$\bigtriangledown\tau_t = \epsilon_t + 0.07\epsilon_{t-1} - 0.30\epsilon_{t-2} - 0.15\epsilon_{t-3} - 0.20\epsilon_{t-4}.$$

The residual variance is 2.475. All these parameters are obtained by maximizing the Gaussian likelihood so that problem (1.2) is of some importance. Shea does not discuss the normality of the innovations in assessing the fit of this model but rather goes on to find a good model for the bivariate series based on an analysis of the residual's cross correlation matrix.

An application of the methods of the paper yields $\mathcal{R}_3 = 22.85$, with a p-value of 0.00004 while $\mathcal{R}_{\hat{K}}(2) = 22.77$ ($\hat{K} = 2$) yielding a p-value of 0.00003 according to (3.3). Thus, both tests strongly reject the null hypothesis (1.2). A complementary analysis helps in understanding what aspect of this null distribution is not supported by the data. We find that $\mathcal{R}_1 = 0.15$ ($p = 0.69$) with a skewness coefficient of 0.13. Thus there is no reason to suspect an asymmetrical distribution for the innovations. On the other hand, we can notice that 9.3% of the absolute residuals are greater than 2.5 and the kurtosis is 4.33. Thus, if the model entertained above is correct, the conclusion that emerges from the present analysis is that the $\bigtriangledown\tau_t$ series could have been generated from innovations with a symmetric distribution having fatter tails than the Gaussian.

ACKNOWLEDGEMENT
The authors would like to thank Dr. B.L. Shea for some insight on subroutine G13DCF of the NAG library and for providing them with the data set used in Section 5.

Appendix A

We show that, for the Legendre polynomial $L_k(.)$ satisfying (2.3), we have under H_0,

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\boldsymbol{\beta}}L_k(U_t)\xrightarrow{P}\begin{bmatrix}\underset{(p+q)\times 1}{\mathbf{0}}\\ -\frac{1}{\sigma}b_k\end{bmatrix}$$

so that the convergence (2.8) holds. Note that it suffices to show

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\sigma}L_k(U_t)\xrightarrow{P}E\left[\frac{\partial}{\partial\sigma}L_k(U_t)\right]=-\frac{1}{\sigma}b_k,\tag{A.1.a}$$

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\varphi_1}L_k(U_t)\xrightarrow{P}E\left[\frac{\partial}{\partial\varphi_1}L_k(U_t)\right]=0,\tag{A.1.b}$$

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta_1}L_k(U_t)\xrightarrow{P}E\left[\frac{\partial}{\partial\theta_1}L_k(U_t)\right]=0.\tag{A.1.c}$$

If either p or $q=0$, some of the convergences above are obvious. Thus, to avoid trivialities, we assume in the sequel that p and $q>0$. First,

$$\frac{\partial}{\partial\sigma}L_k(U_t)=-\frac{2\epsilon_t}{\sigma^2}\phi\left(\frac{\epsilon_t}{\sigma}\right)L_k'(x)|_{x=2\Phi\left(\frac{\epsilon_t}{\sigma}\right)-1}=-\frac{\epsilon_t}{\sigma^2}w\left(\frac{\epsilon_t}{\sigma}\right)\text{ say.}$$

The law of large numbers yields (A.1.a). For (A.1.b), define for $r\geq 0$,

$$B_{r-1}=\frac{\partial}{\partial\varphi_1}\delta_r(\boldsymbol{\theta},\boldsymbol{\varphi}),$$

where, setting $\varphi_0=-1$, $\gamma_0=\theta_0=1$, we have

$$\delta_r(\boldsymbol{\theta},\boldsymbol{\varphi})=\delta_r=\sum_{i=0}^{\min(r,p)}\varphi_i\gamma_{r-i}\quad r\geq 0\tag{A.2}$$

and

$$\gamma_r=-\sum_{i=1}^{\min(r,q)}\gamma_{r-i}\theta_i\quad r\geq 1.\tag{A.3}$$

Obviously $B_{r-1} = \gamma_{r-1}$ when $r \geq 1$. For $r \geq q$, from Brockwell & Davis (1991), p.107,

$$\gamma_r = \sum_{i=1}^{j} \sum_{n=0}^{r_i - 1} c_{in} r^n \alpha_i^r$$

for some constants c_{in} and where the α_i's are the j distinct roots of $1 + \theta_1 z + ... + \theta_q z^q$ and r_i is the multiplicity of α_i, $i = 1, ..., j$. Thus, when $r \geq q + 1$,

$$B_{r-1} = \sum_{i=1}^{j} \sum_{n=0}^{r_i - 1} c_{in} (r-1)^n \alpha_i^{-r+1}. \tag{A.4}$$

It is well known that if $(X_t, t \in \mathbb{Z})$ is a weak stationary process with auto-covariance function $Cov(X_t, X_{t+h})$ that tends to 0 as $h \to \infty$, then $\overline{X}_T \xrightarrow{P} E(X_t)$. We apply this result with $X_t = \partial L_k(U_t)/\partial \varphi_1$ which are obviously identically distributed. From (1.1) and (2.1), we have

$$X_t = \frac{1}{\sigma} w\left(\frac{\epsilon_t}{\sigma}\right) \frac{\partial}{\partial \varphi_1} \epsilon_t = -\frac{1}{\sigma} w\left(\frac{\epsilon_t}{\sigma}\right) \left(Y_{t-1} - \sum_{r=0}^{\infty} \left(\frac{\partial}{\partial \varphi_1} \delta_r\right) \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{Y}_{t-1-r}^{(q)} \right). \tag{A.5}$$

Hence $E(X_t) = 0$. Moreover, $Var(X_t) < \infty$ as shown in Appendix C and, as shown in Appendix D, $Cov(X_t, X_{t+h})$ depends on h and not on t. Thus $(X_t, t \in \mathbb{Z})$ is stationary. We now show that $Cov(X_t, X_{t+h}) \to 0$ as h increases.

From (A.5), we get that, for h large enough, $|Cov(X_t, X_{t+h})| = |d_1| E|w(\epsilon_{t+h}/\sigma)|/\sigma$, where

$$d_1 = E\left[\frac{1}{\sigma} w\left(\frac{\epsilon_t}{\sigma}\right) \left\{ Y_{t-1} - \sum_{r=0}^{\infty} B_{r-1} \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{Y}_{t-1-r}^{(q)} \right\} \left\{ Y_{t+h-1} - \sum_{r=0}^{\infty} B_{r-1} \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{Y}_{t+h-1-r}^{(q)} \right\} \right]. \tag{A.6}$$

But, $|d_1| \leq d_2 + \sum_{j=1}^{q} |\theta_j|(d_{3j} + d_{4j}) + \sum_{i=1}^{q} \sum_{j=1}^{q} |\theta_i \theta_j| d_{5ij}$ where

$$d_2 = \left| E\left[\frac{1}{\sigma} w\left(\frac{\epsilon_t}{\sigma}\right) Y_{t-1} Y_{t+h-1} \right] \right|, \; d_{3j} = \sum_{r=0}^{\infty} \left| B_{r-1} E\left[\frac{1}{\sigma} w\left(\frac{\epsilon_t}{\sigma}\right) Y_{t-r-j} Y_{t+h-1} \right] \right|,$$

$$d_{4j} = \sum_{r=0}^{\infty} \left| B_{r-1} E\left[\frac{1}{\sigma} w\left(\frac{\epsilon_t}{\sigma}\right) Y_{t-1} Y_{t+h-r-j} \right] \right|$$

and

$$d_{5ij} = \sum_{r=0}^{\infty} \sum_{r'=1}^{\infty} \left| B_{r-1} B_{r'-1} E \left[\frac{1}{\sigma} w \left(\frac{\epsilon_t}{\sigma} \right) Y_{t-r-i} Y_{t+h-r'-j} \right] \right|.$$

It can be shown that d_2, d_{3j}, d_{4j} and $d_{5ij} \to 0$ when $h \to \infty$. Proof for d_{4j}, which is typical, is sketched in Appendix E. This yields (A.1.b).

As for (A.1.c), let $A_r = \frac{\partial}{\partial \theta_1} \delta_r(\boldsymbol{\theta}, \boldsymbol{\varphi})$. From (A.2), we have for all $r \geq p$,

$$A_r = \sum_{l=0}^{p} \varphi_l \frac{\partial}{\partial \theta_1} \gamma_{r-l} = \sum_{l=0}^{p} \varphi_l \gamma'_{r-l}.$$

But also, from (A.3), for $r \geq q$,

$$\gamma'_r = -(\theta_1 \gamma'_{r-1} + ... + \theta_q \gamma'_{r-q}) - \gamma_{r-1}.$$

We thus obtain, for $r \geq q$, the system

$$\begin{cases} \gamma'_r + \theta_1 \gamma'_{r-1} + ... + \theta_q \gamma'_{r-q} = -\gamma_{r-1} \\ \gamma_r + \theta_1 \gamma_{r-1} + ... + \theta_q \gamma_{r-q} = 0 \end{cases}$$

from which we find

$$0 = \sum_{j=0}^{q} \theta_j \left(- \sum_{i=0}^{q} \theta_i \gamma'_{r-j-i+1} \right) \text{ for all } r \geq 2q-1.$$

Grouping the terms in γ' with the same value of $j+i = h$ yields, for $r \geq 2q-1$,

$$0 = \sum_{h=0}^{2q} a_h \gamma'_{r-h+1} \text{ where } a_h = \sum_{\substack{i+j=h \\ 0 \leq i,j \leq q}} \theta_i \theta_j.$$

Hence for $r \geq 2q$,

$$\gamma'_r + \sum_{h=1}^{2q} a_h \gamma'_{r-h} = 0.$$

Again from Brockwell & Davis (1991), p.107, we have, for some constants d_{in}

$$\gamma'_r = \sum_{i=1}^{j} \sum_{n=0}^{s_i - 1} d_{in} r^n \beta_i^{-r}$$

where the β_i´s are the j distinct roots of $1 + a_1 z + a_2 z^2 + ... + a_{2q} z^{2q}$ and s_i is the multiplicity of β_i, $i = 1, .., j$. Now

$$\left(\sum_{i=0}^{q} \theta_i z \right)^2 = \sum_{h=0}^{2q} \left(\sum_{\substack{i+j=h \\ 0 \leq i,j \leq q}} \theta_i \theta_j \right) z^h = \sum_{h=0}^{2q} a_h z^h$$

where $a_0 = \theta_0^2 = 1$. This shows that the roots of $1 + a_1 z + a_2 z^2 + ... + a_{2q} z^{2q}$ are exactly the same than that of $1 + \theta_1 z + \theta_2 z^2 + ... + \theta_q z^q$, apart from the multiplicity. Thus, we obtain

$$A_r = \sum_{l=0}^{p} \sum_{i=1}^{j} \sum_{n=0}^{s_i - 1} d_{in} (r - l)^n \alpha_i^{-(r-l)} \varphi_l \tag{A.7}$$

for all $r \geq \max(2q, p)$. By the same argument as previously, but using expression (A.7) of A_r instead of B_{r-1}, we get (A.1.c).

Appendix B

We show that $E(\boldsymbol{V}_t) = 0$ and $Var(\boldsymbol{B}\boldsymbol{V}_t) = \boldsymbol{I}_K - \boldsymbol{b}_K \boldsymbol{b}_K^{\mathsf{T}}/2$. In view of (1.1) and (2.1),

$$\frac{\partial}{\partial \boldsymbol{\varphi}} Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right) = \frac{\epsilon_t}{\sigma^2}\left[\boldsymbol{Y}_{t-1}^{(p)} - \sum_{r=0}^{\infty}\left(\frac{\partial}{\partial \boldsymbol{\varphi}}\delta_r\right)\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{Y}_{t-1-r}^{(q)}\right],$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right) = \frac{\epsilon_t}{\sigma^2}\left[\boldsymbol{\epsilon}_{t-1}^{(q)} - \sum_{r=0}^{\infty}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\delta_r\right)\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{Y}_{t-1-r}^{(q)}\right]$$

and

$$\frac{\partial}{\partial \sigma} Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right) = \frac{1}{\sigma}\left(\left(\frac{\epsilon_t}{\sigma}\right)^2 - 1\right).$$

From these expressions, it follows that $E(\boldsymbol{V}_t) = 0$ under H_0. Let $\boldsymbol{L}_t = (L_1(U_t), ..., L_K(U_t))^{\mathsf{T}}$. We have, again under H_0, $Var(\boldsymbol{L}_t) = \boldsymbol{I}_K$. Thus,

$$Cov\left(\boldsymbol{L}_t, \frac{\partial}{\partial \boldsymbol{\beta}} Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right)\right)^{\mathsf{T}} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \frac{1}{\sigma}\boldsymbol{b}_K^{\mathsf{T}} \end{bmatrix} = \mathcal{J}_K^{\mathsf{T}}.$$

Finally, $Var\left(\frac{\partial}{\partial \boldsymbol{\beta}} Log\left(\frac{1}{\sigma}\phi\left(\frac{\epsilon_t}{\sigma}\right)\right)\right) = \mathcal{I}_\beta = \begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \frac{2}{\sigma^2} \end{bmatrix}$, for some matrix $\boldsymbol{C}$ whose exact expression will not be needed. Thus

$$Var(\boldsymbol{V}_t) = \begin{bmatrix} \boldsymbol{I}_K & \mathcal{I}_K \\ \mathcal{I}_K^{\mathsf{T}} & \mathcal{I}_\beta \end{bmatrix},$$

which yields the desired result.

Appendix C

We show that $Var(X_t) < \infty$. Without loss of generality, set $\sigma = 1$. This will be assumed here and throughout the following appendices. Since Y_t is causal, we can write $Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ and from (A.5)

$$Var(X_t) = E(w(\epsilon_t))^2 E\left(Y_{t-1} - \sum_{r=0}^{\infty} B_{r-1} \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{Y}_{t-1-r}^{(q)}\right)^2 = E(w(\epsilon_t))^2 E\left(\sum_{h=1}^{\infty} d_h \epsilon_{t-h}\right)^2$$

where $d_h = \psi_{h-1} - \sum_{\substack{r+j+l=h \\ 1\le j\le q \\ 0\le r,l\le h-1}} \psi_l \gamma_{r-1} \theta_j$. We now need the following lemma.

Lemma C.1. *If the ARMA process (1.1) is causal and invertible, then* $\sum_{h=1}^{\infty} |d_h| < \infty$.

Proof. From (A.3), $\sum_{\substack{r+j+l=h \\ 1\le j\le q \\ 0\le r,l\le h-1}} \psi_l \gamma_{r-1} \theta_j = \sum_{k=1}^{h} \psi_{h-k} \left(\sum_{\substack{r+j=k \\ 1\le j\le q \\ 0\le r\le h-1}} \gamma_{r-1} \theta_j \right) = -\sum_{k=1}^{h} \psi_{h-k} \gamma_{k-1}$. We have also $\sum_{h=1}^{\infty} \sum_{k=1}^{h} |\psi_{h-k} \gamma_{k-1}| = \sum_{h=0}^{\infty} \sum_{k=0}^{h} |\psi_{h-k} \gamma_k| = \sum_{k=0}^{\infty} |\gamma_k| \sum_{h=0}^{\infty} |\psi_h|$. Thus, $\sum_{h=1}^{\infty} |d_h| \le \sum_{h=1}^{\infty} |\psi_{h-1}| + \sum_{h=1}^{\infty} \sum_{k=1}^{h} |\psi_{h-k} \gamma_{k-1}| = \sum_{h=0}^{\infty} |\psi_h| \left(\sum_{k=0}^{\infty} |\gamma_k| + 1\right)$. But from Brockwell & Davis (1991), p.87, $\sum_{k=0}^{\infty} |\gamma_k|$ is finite. Since under the assumption of Theorem 2.1, $\sum_{j=0}^{\infty} |\psi_j| < \infty$ the lemma follows. $\square$

From this lemma, we conclude that $E\left[\sum_{h=1}^{\infty} d_h \epsilon_{t-h}\right]^2 = \sum_{h=1}^{\infty} d_h^2 < \infty$. Since

$$E(w(\epsilon_t))^2 = 4 \int (L_k'(2\Phi(x) - 1))^2 \phi^3(x) dx < \infty,$$

the result follows.

Appendix D

Here, we show that $Cov(X_t, X_{t+h})$ depends on h and not on t. We have

$$
\begin{aligned}
Cov(X_t, X_{t+h}) = E(X_t X_{t+h}) &= E\left(w(\epsilon_t) \sum_{l=1}^{\infty} d_l \epsilon_{t-l} w(\epsilon_{t+h}) \sum_{l=1}^{\infty} d_l \epsilon_{t+h-l} \right) \\
&= E(w(\epsilon_{t+h}))E(w(\epsilon_t))E\left(\sum_{l=1}^{\infty} d_l \epsilon_{t-l} \right) E\left(\sum_{l=1}^{h-1} d_l \epsilon_{t+h-l} \right) \\
&\quad + d_h E(w(\epsilon_{t+h}))E(w(\epsilon_t)\epsilon_t)E\left(\sum_{l=1}^{\infty} d_l \epsilon_{t-l} \right) \\
&\quad + E(w(\epsilon_{t+h}))E(w(\epsilon_t))E\left(\sum_{l=1}^{\infty} d_l \epsilon_{t-l} \sum_{l=1}^{\infty} d_{l+h} \epsilon_{t-l} \right) \\
&= d_h c_1 E\left(\sum_{l=1}^{\infty} d_l \epsilon_{t-l} \right) + E^2(w(\epsilon_t))E\left(\sum_{l=1}^{\infty} d_l d_{l+h} \epsilon_{t-l}^2 \right) \text{ say.} \quad \text{(D.1)}
\end{aligned}
$$

The first term od (D.1) vanishes while by Lemma C.1 and Schwarz´s inequality, the second term is finite. Thus

$$
Cov(X_t, X_{t+h}) = E^2(w(\epsilon_t)) \sum_{l=1}^{\infty} d_l d_{l+h}.
$$

Appendix E

Here we sketch the proof that the typical element d_{4j} of inequality (A.6) vanishes. From $Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ and the fact that the remainder of a convergent series converges toward 0, we have

$$\lim_{h\to\infty} d_{4j} = \lim_{h\to\infty} \sum_{r=0}^{\infty} \left| B_{r-1} E\left(w(\epsilon_t) Y_{t-1} Y_{t+h-r-j}\right) \right| \leq \lim_{h\to\infty} |E(w(\epsilon_t))| \sum_{r=0}^{h-j} \left| B_{r-1} \sum_{a=0}^{\infty} \psi_a \psi_{a+h-j-r+1} \right|$$

$$\leq |E(w(\epsilon_t))| \lim_{h\to\infty} \left[\sum_{a=0}^{m-1} |\psi_a| \sum_{r=0}^{a+h-j-m} |B_{r-1}\psi_{a+h-j-r+1}| + \sum_{r=a+h-j-m+1}^{h-j} |B_{r-1}\psi_{a+h-j-r+1}| \right.$$
$$\left. + \sum_{r=0}^{h-j} |B_{r-1}| \sum_{a=m}^{\infty} |\psi_a \psi_{a+h-j-r+1}| \right] \quad \text{(E.1)}$$

where $m = \max\{p, q+1\} - p$. Now, for the first term in the limit of (E.1), we have, using the expression for B_{r-1} given in (A.4) and that of $\psi_{a+h-j-r+1}$ given in Brockwell & Davis (1991) eq. (3.3.6)

$$\sum_{r=q+1}^{a+h-j-m} |B_{r-1}\psi_{a+h-j-r+1}| =$$
$$\sum_{r=q+1}^{a+h-j-m} \left| \sum_{b=1}^{k} \sum_{l=0}^{r_b-1} c_{bl} r^l \alpha_a^{-r} \sum_{b'=1}^{k'} \sum_{l'=0}^{r_{b'}-1} \alpha_{b'l'} (a+h-j-r+1)^{l'} \xi_{b'}^{-(a+h-j-r+1)} \right|$$
$$\leq \sum_{b=1}^{k} \sum_{l=0}^{r_b-1} \sum_{b'=1}^{k'} \sum_{l'=0}^{r_{b'}-1} \sum_{d=0}^{l'} \binom{l'}{d} \left\{ |c_{bl}\alpha_{b'l'}| |\xi_{b'}^{-(a+h-j+1)} (a+h-j+1)^{l'-d}| \sum_{r=q+1}^{a+h-j-m} r^{l+d} |\alpha_a|^{-r} |\xi_{b'}|^r \right\}.$$

If $|\xi_{b'}| < |\alpha_a|$, the last sum is finite (by D´Alambert´s rule) and, obviously, the corresponding term in braces converges toward 0 as $h \to \infty$. So suppose that $|\alpha_a| = 1 + \epsilon_1 < |\xi_{b'}| = 1 + \epsilon_2$ with ϵ_1 and ϵ_2 positive. We show that this term in braces still converges towards 0. Now,

$$\left| \frac{(a+h-j+1)^{l'-d}}{\xi_{b'}^{(a+h-j+1)}} \right| \sum_{r=q+1}^{a+h-j-m} r^{l+d} |\alpha_a|^{-r} |\xi_{b'}|^r \leq \frac{|a+h-j+1|^{l'-d}}{|\xi_{b'}|^{a+h-j+1}} \sum_{r=0}^{a+h-j+1} r^{l+d} \left(\frac{|\xi_{b'}|}{|\alpha_a|} \right)^r . \quad \text{(E.2)}$$

But, for all $\epsilon > 0$, there exist a positive constant C such that

$$\sum_{r=0}^{a+h-j+1} r^{l+d} \left(\frac{|\xi_{b'}|}{|\alpha_a|}\right)^r \leq C \sum_{r=0}^{a+h-j+1} \left(\frac{|\xi_{b'}|}{|\alpha_a|} + \epsilon\right)^r .$$

Hence, the left-hand side of (E.2) is bouded above by

$$C\frac{|a+h-j+1|^{l'-d}}{|\xi_{b'}|^{a+h-j+1}} \frac{\left(\frac{|\xi_{b'}|}{|\alpha_a|} + \epsilon\right)^{h+a-j+2} - 1}{\frac{|\xi_{b'}|}{|\alpha_a|} + \epsilon - 1} \leq C'|a+h-j+1|^{l'-d} \left(\frac{\frac{|\xi_{b'}|}{|\alpha_a|} + \epsilon}{|\xi_{b'}|}\right)^{a+h-j+2} \tag{E.3}$$

for some $C' > 0$. If, in (E.3), one takes $\epsilon > 0$ but smaller than $\epsilon_1(1+\epsilon_2)/(1+\epsilon_1)$, then

$$\frac{|\xi_{b'}|}{|\alpha_a|} + \epsilon - |\xi_{b'}| = \frac{-\epsilon_1(1+\epsilon_2)}{1+\epsilon_1} + \epsilon < 0$$

and the right hand side of (E.3) converges to 0 as $h \to \infty$. This shows that the first term in the limit of (E.1) converges to 0. It follows that the second term also converges toward 0. As for the last term in the limit, a similar argument yields

$$\sum_{r=0}^{h-j} |B_{r-1}| \sum_{a=m}^{\infty} |\psi_a \psi_{a+h-j-r+1}| \leq$$
$$\sum_{b=1}^{k} \sum_{l=0}^{r_b-1} \sum_{b'=1}^{k'} \sum_{l'=0}^{r_{b'}-1} \sum_{d=0}^{l'} \binom{l'}{d} \left| \alpha_{bl}\alpha_{b'l'} \frac{\beta_{l,d,b,b'}}{\xi_{b'}^{(-j+1)}} \right| \sum_{r=0}^{h-j} \left| B_{r-1} \frac{(h-j-r+1)^{l'-d}}{\xi_{b'}^{(h-r)}} \right|,$$

where $\beta_{l,d,b,b'} = \sum_{a=m}^{\infty} a^{l+d}|\xi_b \xi_{b'}|^{-a} < \infty$. Now

$$\sum_{r=0}^{h-j} \left| B_{r-1} \frac{(h-j-r+1)^{l'-d}}{\xi_{b'}^{(h-r)}} \right| \leq \sum_{r=0}^{q} \left| B_{r-1} \frac{(h-j-r+1)^{l'-d}}{\xi_{b'}^{(h-r)}} \right| + \sum_{r=q+1}^{\infty} \left| B_{r-1} \frac{(h-j-r+1)^{l'-d}}{\xi_{b'}^{(h-r)}} \right|.$$

The first term on the right hand side of this expression converges to 0. As for the second term,

$$\sum_{r=q+1}^{\infty} \left| B_{r-1} \xi_{b'}^{-(h-r)} (h-j-r+1)^{l'-d} \right| \leq$$
$$\sum_{e=0}^{l'-d} \sum_{u=1}^{k} \sum_{\nu=0}^{r_u-1} \binom{l'-d}{e} |c_{uv}||\xi_{b'}^{-h}(h-j+1)^{l'-d-e}| \sum_{r=q+1}^{\infty} r^{\nu+e} \left(\frac{|\xi_{b'}|}{|\alpha_u|}\right)^r .$$

By the same argument as in (E.2), this term converges to 0 as $h \to \infty$. Thus all terms on the right hand side of (E.1) converge to 0 so that $d_{4j} \to 0$.

References

ANDĚL, J. (1997). On residual analysis for time series models. *Kybernetika* **33**, 161–170.

BEISER, A. (1985). *Distributions of $\sqrt{b_1}$ and b_2 for autoregressive errors.* Ph.D. thesis, Boston University.

BROCKETT, P. L., HINICH, M. J. & PATTERSON, D. (1988). Bispectral-based tests for the detection of gaussianity and linearity in time series. *J. Amer. Statist. Assoc.* **83**, 657–664.

BROCKWELL, P. J. & DAVIS, R. A. (1991). *Time series: Theory and Methods.* Springer-Verlag New York, 2nd ed.

BURN, D. (1987). Simulation of stationary time series. *Proceedings of the 1987 Winter Simulation Conference* , 289–294.

CROMWELL, J. B., LABYS, W. C. & TERRAZA, M. (1994). *Univariate tests for time-series models.* Sage Publications Inc, Thousand Oaks, California.

D'AGOSTINO, R. B. & STEPHENS, M. A. (1986). *Goodness-of-fit techniques.* Statistics: TextBOOKs and Monographs, 68, New York: Marcel Dekker.

DOORNIK, J. & HANSEN, H. (1994). An omnibus test for univariate and multivariate normality. Tech. rep., Nuffiled College, Oxford.

EPPS, T. W. (1987). Testing that a stationnary time series is gaussian. *Ann. Statist.* **15**, 1683–1698.

FRANCES, P. (1998). *Time series models for business and economic forecasting.* Cambridge University Press, Cambridge.

GASSER, T. (1975). Goodness-of-fit tests for correlated data. *Biometrika* **62**, 563–570.

GOURIÉROUX, C. & MONFORT, A. (1995). *Séries temporelles et modèles dynamiques.* Economica, 2nd ed.

GRANGER, C. W. J. (1976). Tendency towards normality of linear combinations of random variables. *Metrika* **23**, 237–248.

HENZE, N. (1997). Do components of smooth tests of fit have diagnostic properties? *Metrika* **45**, 121–130.

HEUTS, R. & RENS, S. (1986). Testing normality when observations satisfy a certain low order arma-scheme. *Computat. Statist. Quarterly* **1**, 49–60.

HINICH, M. J. (1982). Testing for gaussianity and linearity of a stationary time series. *J. Time Ser. Anal.* **3**, 169–176.

HIPEL, K. W. & MCLEOD, A. I. (1994). *Time series modelling of water resources and environmental systems.* [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam] (New York; Amsterdam).

JARQUE, C. & BERA, A. (1987). A test for normality of observations and regression residuals. *Internat. Statist. Review* **55**, 163–172.

KALLENBERG, W. & LEDWINA, T. (1997a). Data driven smooth tests for composite hypotheses: comparison of powers. *J. Statist. Comput. Simul.* **59**, 101–121.

KALLENBERG, W. & LEDWINA, T. (1997b). Data-driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.* **92**, 1094–1104.

KOUL, H. L. & STUTE, W. (1999). Nonparametric model checks for time series. *Ann. Statist.* **27**, 204–236.

LEDWINA, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89**, 1000–1005.

LEE, S. & NA, S. (2002). On the Bickel-Rosenblatt test for first-order autoregressive models. *Statist. Probab. Lett.* **56**, 23–35.

LOMNICKI, Z. (1961). Tests for departure from normality in the case of linear stochastic processes. *Metrika* **4**, 37–62.

LUTKEPOHL, H. & SCHNEIDER, W. (1989). Testing for normality of autoregressive time series. *Comput. Statist. Quaterly* **2**, 151–168.

MOORE, D. S. (1982). The effect of dependence on chi squared tests of fit. *Ann. Statist.* **10**, 1163–1171.

Neyman, J. (1937). Smooth test for goodness of fit. *Skand. Aktuar.* **20**, 149–199.

Ojeda, R., Cardoso, J. & Moulines, E. (1997). Asymptotically invariant gaussianity test for causal invertible time series. *Proc. of IEEE international conference on Acoustics, Speech and Signal Processing* **5**, 3713–3716.

Pierce, D. A. & Gray, R. J. (1985). Goodness-of-fit tests for censored survival data. *Ann. Statist.* **13**, 552–563.

Sansone, G. (1959). *Orthogonal functions.* New York: Interscience.

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

Shea, B. L. (1987). Estimation of multivariate time series. *J. Time Ser. Anal.* **8**, 95–109.

Thomas, D. R. & Pierce, D. A. (1979). Neyman's smooth goodnes-of-fit test when the hypothesis is composite. *J. Amer. Statist. Assoc.* **74**, 441–445.

Weisberg, S. & Bingham, C. (1975). An approximate analysis of variance test for non-normality suitable for machine calculation. *Technometrics* **17**, 133–134.

Table 4.1.a: Distribution (in % of the number of parameter points) of the empirical p-values (based on 10000 replications) for the test based on $\mathcal{R}_3$ among 5 sub-intervals.

$\mathcal{R}_3$			Observed level					Min
Model	T	α	I_1	I_2	I_3	I_4	I_5	p-level
MA(2) (64 points)	50	5%	18.8	68.8	12.5	0	0	2.76
	100	5%	1.6	50.0	48.4	0	0	3.41
	200	5%	0	9.4	89.1	1.6	0	4.04
	50	10%	23.4	53.1	23.4	0	0	6.49
	100	10%	6.3	20.3	73.4	0	0	7.96
	200	10%	0	7.8	90.6	1.6	0	8.92
AR(2) (64 points)	50	5%	17.2	82.8	0	0	0	3.13
	100	5%	3.1	45.3	51.6	0	0	3.48
	200	5%	0	12.5	87.5	0	0	4.07
	50	10%	28.1	59.4	12.5	0	0	7.09
	100	10%	12.5	18.8	68.8	0	0	7.29
	200	10%	1.6	7.8	90.6	0	0	8.41
ARMA(1,2) (250 points)	50	5%	47.2	46.4	6.4	0	0	2.43
	100	5%	8.0	71.6	20.4	0	0	2.98
	200	5%	0.8	32.4	66.4	0.4	0	3.32
	50	10%	65.6	24.0	10.4	0	0	6.20
	100	10%	21.6	35.2	42.8	0.4	0	6.80
	200	10%	4.0	19.6	75.6	0.8	0	7.42
ARMA(2,1) (250 points)	50	5%	28.0	68.0	4.0	0	0	2.81
	100	5%	5.6	54.0	38.0	2.0	0.4	3.19
	200	5%	0	19.2	78.8	2.0	0.4	3.19
	50	10%	42.2	45.2	12.4	0	0	6.71
	100	10%	20.0	17.2	61.2	1.2	0.4	6.97
	200	10%	0.8	19.2	78.8	1.2	0	8.08
ARMA(2,2) (294 points)	50	5%	41.2	57.1	1.7	0	0	2.56
	100	5%	5.1	74.1	20.8	0	0	3.09
	200	5%	0.3	27.9	71.8	0	0	3.47
	50	10%	57.8	37.4	4.8	0	0	6.24
	100	10%	21.1	33.7	45.2	0	0	6.88
	200	10%	3.1	18.0	78.6	0.3	0	7.86

Table 4.1.b: Distribution (in % of the number of parameter points) of the empirical p-values (based on 10000 replications) for the test based on $\mathcal{R}_{\hat{K}}(2)$ among 5 sub-intervals.

$\mathcal{R}_{\hat{K}}(2)$			Observed level					Min
Model	T	α	I_1	I_2	I_3	I_4	I_5	p-level
	50	5%	0	9.4	46.9	43.7	0	4.12
	100	5%	0	14.1	68.8	17.2	0	4.17
MA(2)	200	5%	0	7.8	87.5	4.7	0	4.23
	50	10%	6.3	14.1	62.5	17.2	0	7.78
(64 points)	100	10%	6.3	6.3	81.3	6.3	0	8.19
	200	10%	0	6.3	89.1	4.7	0	8.83
	50	5%	0	10.9	53.1	35.9	0	3.98
	100	5%	0	15.6	59.4	25.0	0	4.11
AR(2)	200	5%	0	4.7	67.2	28.1	0	4.34
	50	10%	9.4	15.6	70.3	4.7	0	7.91
(64 points)	100	10%	6.3	7.8	85.9	0	0	8.11
	200	10%	1.6	4.7	93.8	0	0	8.43
	50	5%	0	38.8	46.8	14.4	0	3.53
	100	5%	0	34.8	59.2	6.0	0	3.74
ARMA(1,2)	200	5%	0	23.2	75.2	1.6	0	3.80
	50	10%	24.8	31.6	35.2	8.4	0	7.06
(250 points)	100	10%	13.2	27.2	57.6	2.0	0	7.33
	200	10%	4.4	18.0	76.0	1.6	0	7.61
	50	5%	0	20.0	62.0	17.6	0.4	3.80
	100	5%	0	26.4	50.0	22.4	1.2	3.89
ARMA(2,1)	200	5%	0	10.8	68.4	20.4	0.4	4.11
	50	10%	16.0	21.2	54.4	8.4	0	7.17
(250 points)	100	10%	10.4	13.2	66.0	9.2	1.2	7.78
	200	10%	1.2	13.2	81.6	4.0	0	7.91
	50	5%	0	32.0	55.4	12.6	0	3.65
	100	5%	0	31.0	62.9	6.1	0	3.75
ARMA(2,2)	200	5%	0	23.8	75.9	0.3	0	3.89
	50	10%	21.4	30.3	47.0	1.4	0	7.14
(294 points)	100	10%	11.2	24.8	62.6	1.4	0	7.51
	200	10%	2.7	16.7	80.6	0	0	7.62

Table 4.2: Distribution (in % of the number of parameter points) of the empirical p-values (based on 10000 replications) of various tests for the ARMA(1,2) model. $\mathcal{AD}$=Anderson-Darling, $\mathcal{WB}$=Weisberg-Bingham, $\mathcal{JB}$=Jarque-Bera and $\mathcal{R}_{\hat{K}}(1) = \mathcal{R}_{\hat{K}}$ with $d = 1$.

Test			Observed level					Min
Model	T	α	I_1	I_2	I_3	I_4	I_5	p-level
$\mathcal{AD}$	50	5%	43.8	24.4	20.0	6.8	0	0.54
	100	5%	32.0	34.8	33.2	0	0	0.92
	200	5%	11.6	38.0	49.6	0.8	0	1.50
	50	10%	41.2	16.0	22.8	18.4	1.6	3.38
	100	10%	23.2	24.4	48.0	3.6	0.8	3.93
	200	10%	9.6	13.2	70.0	6.8	0.4	4.65
$\mathcal{WB}$	50	5%	61.2	23.6	15.2	0	0	0.57
	100	5%	39.2	46.0	14.4	0.4	0	0.93
	200	5%	10.8	30.4	56.0	2.8	0	1.60
	50	10%	56.8	16.4	25.6	1.2	0	2.96
	100	10%	37.2	47.2	15.6	0	0	3.55
	200	10%	15.6	36.0	46.0	2.4	0	4.71
$\mathcal{JB}$	50	5%	71.2	28.8	0	0	0	3.13
	100	5%	0.4	99.2	0.4	0	0	3.13
	200	5%	0	85.2	14.4	0.4	0	4.18
	50	10%	100	0	0	0	0	4.96
	100	10%	100	0	0	0	0	5.88
	200	10%	98.8	1.2	0	0	0	7.23
$\mathcal{R}_{\hat{K}}(1)$	50	5%	0	58.0	27.6	14.4	0	3.52
	100	5%	10.0	48.8	39.2	2.0	0	3.32
	200	5%	8.4	33.6	56.4	1.6	0	3.39
	50	10%	43.6	20.8	26.0	9.6	0	4.69
	100	10%	26.0	24.8	48.0	1.2	0	4.81
	200	10%	8.8	16.4	70.4	4.4	0	5.79

Table 4.3.a: Empirical power (based on 10000 replications with $\alpha = 10\%$) of various tests for the ARMA(2,1) model with parameter $(\boldsymbol{\varphi}, \theta_1) = (-0.8, -0.1, 0.7)$. The part above the line in the middle of the table corresponds to symmetric alternatives while those below are skewed. The distributions are ordered according to increasing kurtosis.

	$T = 50$				$T = 100$			
Alternatives	$\mathcal{R}_3$	$\mathcal{R}_{\hat{K}}(2)$	$\mathcal{AD}$	$\mathcal{WB}$	$\mathcal{R}_3$	$\mathcal{R}_{\hat{K}}(2)$	$\mathcal{AD}$	$\mathcal{WB}$
SB(0;0.5)	83.19	84.76	33.12	28.93	99.72	99.84	78.34	86.81
TU(1.5)	66.94	67.96	23.18	18.27	97.50	98.14	52.60	64.10
TU(0.7)	44.47	45.64	17.02	12.42	85.82	87.84	29.28	32.21
Logistic(1)	20.74	22.75	10.70	19.02	32.94	34.29	12.78	27.41
TU(10)	94.64	96.64	78.08	83.60	99.95	99.99	99.50	99.51
SC(0.05;3)	33.65	37.38	15.46	35.98	50.79	55.71	22.44	56.32
SC(0.2;5)	96.36	96.77	74.67	92.84	99.93	99.94	98.47	99.88
SC(0.05;5)	62.33	65.22	39.56	63.63	84.25	86.38	63.31	86.95
SC(0.05;7)	74.05	76.12	58.22	75.32	92.61	93.71	82.98	93.81
SU(0;1)	75.96	76.49	39.99	66.57	95.88	95.98	73.82	91.12
SB(0.533;0.5)	91.09	89.76	52.47	59.41	99.93	99.85	95.83	99.08
SB(1;1)	53.75	56.94	26.20	32.60	96.93	87.75	51.77	80.02
LC(0.2;3)	55.58	57.79	25.79	29.72	88.79	88.81	52.04	65.87
Weibull(2)	28.10	30.72	16.86	21.25	48.17	50.48	25.24	44.12
LC(0.1;3)	44.10	43.85	21.27	35.21	75.16	72.10	40.61	67.59
χ^2 (df.=10)	41.41	45.84	20.59	34.72	68.55	70.86	38.20	69.71
LC(0.05;3)	29.50	31.25	15.04	28.28	50.57	51.01	23.11	51.66
LC(0.1;5)	96.10	96.00	80.12	95.07	99.98	99.97	99.76	99.98
SU(-1;2)	37.88	38.92	19.15	33.93	61.39	59.93	30.47	57.73
χ^2 (df.=4)	76.13	80.54	42.57	69.78	96.91	98.03	82.09	98.02
LC(0.05;5)	81.48	83.98	51.98	84.41	97.28	97.98	89.09	98.90
LC(0.05;7)	94.26	94.74	88.61	96.28	99.73	99.75	99.45	99.83
SU(1;1)	96.26	96.15	85.74	93.98	99.97	99.93	99.75	99.99
LN(0;1)	99.52	99.68	96.55	99.24	100	100	100	100

Table 4.3.b: Empirical power (based on 10000 replications with $\alpha = 10\%$) of various tests for the ARMA(1,2) model with parameter $(\varphi_1, \boldsymbol{\theta}) = (-0.7, 0.4, 0.5)$. The part above the line in the middle of the table corresponds to symmetric alternatives while those below are skewed. The distributions are ordered according to increasing kurtosis.

	$T = 50$				$T = 100$			
Alternatives	$\mathcal{R}_3$	$\mathcal{R}_{\hat{K}}(2)$	$\mathcal{AD}$	$\mathcal{WB}$	$\mathcal{R}_3$	$\mathcal{R}_{\hat{K}}(2)$	$\mathcal{AD}$	$\mathcal{WB}$
SB(0;0.5)	73.16	74.90	27.25	22.47	99.22	99.41	73.67	82.00
TU(1.5)	57.86	59.43	19.96	15.07	96.24	97.21	47.58	57.86
TU(0.7)	38.69	39.44	15.13	11.02	82.27	84.47	26.69	28.99
Logistic(1)	18.97	20.87	9.83	17.10	31.68	33.36	12.39	26.74
TU(10)	89.57	91.62	65.83	74.31	99.82	99.88	98.78	99.06
SC(0.05;3)	32.82	36.65	14.76	34.40	50.73	55.27	21.30	55.57
SC(0.2;5)	94.21	94.72	66.49	89.12	99.96	99.95	97.87	99.80
SC(0.05;5)	61.43	63.86	37.95	61.81	83.95	86.17	62.27	86.64
SC(0.05;7)	73.00	75.22	56.46	73.89	92.51	93.77	82.78	94.00
SU(0;1)	71.73	72.44	36.03	62.20	95.46	95.73	71.46	90.44
SB(0.533;0.5)	83.62	82.09	44.34	48.17	99.68	99.51	93.60	97.81
SB(1;1)	45.94	48.41	22.54	26.18	83.93	85.13	48.60	75.28
LC(0.2;3)	49.58	52.11	22.92	26.19	86.92	86.98	48.98	62.54
Weibull(2)	25.63	28.66	15.01	18.52	47.28	48.68	23.95	42.00
LC(0.1;3)	40.62	40.54	18.12	31.38	72.04	68.76	37.70	64.91
χ^2 (df.=10)	37.80	41.98	19.09	31.09	66.48	69.08	36.30	66.96
LC(0.05;3)	28.05	29.86	13.31	26.18	48.50	48.59	21.21	49.12
LC(0.1;5)	93.40	93.03	73.06	90.44	99.97	99.99	99.53	99.98
SU(-1;2)	34.38	36.29	17.25	31.12	60.23	58.90	29.04	56.47
χ^2 (df.=4)	71.19	75.76	37.61	63.30	96.17	97.69	79.72	97.22
LC(0.05;5)	78.07	80.38	48.09	80.80	96.81	96.62	87.32	98.53
LC(0.05;7)	94.05	94.73	85.42	96.39	99.80	99.87	99.71	99.97
SU(1;1)	94.18	94.17	80.95	91.39	99.97	99.94	99.66	99.98
LN(0;1)	98.74	98.85	93.54	98.02	100	100	99.98	100

Table 4.3.c: Empirical power (based on 10000 replications with $\alpha = 10\%$) of various tests for the ARMA(2,2) model with parameter $(\boldsymbol{\varphi}, \boldsymbol{\theta}) = (-1.05, -0.4, 0.15, 0.85)$. The part above the line in the middle of the table corresponds to symmetric alternatives while those below are skewed. The distributions are ordered according to increasing kurtosis.

	$T = 50$				$T = 100$			
Alternatives	$\mathcal{R}_3$	$\mathcal{R}_{\hat{K}}(2)$	$\mathcal{AD}$	$\mathcal{WB}$	$\mathcal{R}_3$	$\mathcal{R}_{\hat{K}}(2)$	$\mathcal{AD}$	$\mathcal{WB}$
SB(0;0.5)	63.85	65.44	23.06	17.57	98.01	98.56	61.71	68.27
TU(1.5)	49.17	50.50	18.61	13.62	92.73	94.17	39.50	44.52
TU(0.7)	32.58	33.88	15.31	10.76	76.34	78.93	24.22	23.84
Logistic(1)	18.59	20.36	10.78	17.60	29.74	31.71	11.70	25.66
TU(10)	85.14	87.26	53.85	65.78	99.79	99.90	96.87	97.87
SC(0.05;3)	30.81	34.69	13.65	32.37	48.01	52.64	20.42	52.73
SC(0.2;5)	92.28	92.96	60.49	85.62	99.88	99.88	96.37	99.63
SC(0.05;5)	58.90	62.20	35.03	59.77	82.67	85.03	59.23	85.24
SC(0.05;7)	72.28	74.04	53.48	72.50	92.23	93.38	80.83	93.51
SU(0;1)	68.60	69.79	34.02	59.99	94.26	94.65	68.14	88.84
SB(0.533;0.5)	76.29	74.04	36.26	36.87	99.33	99.03	86.93	94.72
SB(1;1)	42.03	44.17	21.29	23.84	80.01	81.30	43.60	68.22
LC(0.2;3)	43.88	46.16	20.64	22.62	82.56	82.29	43.62	55.83
Weibull(2)	24.32	26.85	15.31	18.68	43.50	45.02	22.21	38.61
LC(0.1;3)	37.00	37.25	17.67	27.97	68.81	65.56	34.60	60.28
χ^2 (df.=10)	35.04	38.91	18.69	29.66	62.38	64.41	32.68	61.22
LC(0.05;3)	25.91	28.32	13.23	24.17	45.26	45.57	20.13	46.12
LC(0.1;5)	91.04	89.90	65.32	86.06	99.95	99.87	98.96	99.89
SU(-1;2)	33.54	35.40	17.11	29.87	57.53	56.60	27.97	54.25
χ^2 (df.=4)	65.96	70.30	34.83	57.70	94.37	95.97	73.18	94.99
LC(0.05;5)	75.48	77.71	43.10	77.58	96.79	97.47	84.56	98.30
LC(0.05;7)	93.44	94.24	82.38	95.88	99.90	99.95	99.62	99.97
SU(1;1)	93.14	93.17	77.94	90.14	99.93	99.87	98.80	99.81
LN(0;1)	97.89	98.16	89.78	96.83	99.99	99.99	99.90	99.99