

The empirical process of autoregressive residuals

Eric Engler and Bent Nielsen
Department of Economics, University of Oxford.

Mail: Nuffield College, Oxford OX1 1NF, UK

E-mail: bent.nielsen@nuf.ox.ac.uk

17 January 2007

Abstract: The empirical process of the residuals from general autoregressions is investigated. If an intercept is included in the regression, the empirical process is asymptotically Gaussian and free of nuisance parameters. This contrasts the known result that in the unit root case without intercept the empirical process is asymptotically non-Gaussian. The result is used to establish asymptotic theory for the Kolmogorov-Smirnov test, Probability-Probability plots, and Quantile-Quantile plots. The link between sample moments and the empirical process of the residuals is established and used to establish the properties of the cumulant based tests for normality referred to as the Jarque-Bera test.

Keywords: Autoregression, Empirical process, Kolmogorov-Smirnov test, Probability-Probability plots, Quantile-Quantile plots, Test for normality.

1 Introduction

In this paper, the asymptotic theory of the empirical process of autoregressive residuals is analysed. This can then be used to establish formal tests on the distribution of the autoregressive innovations using a Kolmogorov-Smirnov test, an Anderson-Darling-type test, a Probability-Probability plot or a Quantile-Quantile plot with confidence bands, or the standard cumulant based tests for normality, known as the Jarque-Bera tests. The empirical process is analysed for a general class of autoregressions, including vector autoregressions and autoregressive distributed lags models. These can have stationary roots, unit roots as well as explosive roots, and include deterministic terms. It is found to be asymptotically Gaussian and free of nuisance parameters as long as an intercept is included in the model. This is important in applications, as the question of innovation distribution can be addressed without having to locate the characteristic roots. Similar result apply for the likelihood-based tests for unit roots and the order of the autoregression, but not for the usual correlograms, see Nielsen (2001, 2006, 2007).

For the basic first order autoregression without deterministic regressors, the asymptotic theory is available in the literature. The results are somewhat discouraging in that the empirical process is asymptotically Gaussian in the stationary and the explosive cases, but non-Gaussian in the unit root case. As a consequence, inference about the distribution of the innovations and about the autoregressive coefficient has to be conducted simultaneously. The key to the results presented

here is that the non-Gaussian component in the unit root case stems from the sum of the residuals. When an intercept is included in the model, as in most applications, this sum is zero due to the estimating equation for the intercept and therefore the non-Gaussian component does not arise.

We will exploit that the expectation taken with respect to the empirical distribution of residuals $\widehat{\varepsilon}_1, \dots, \widehat{\varepsilon}_T$ is simply their sample average. This is first used as an expository device, and is later used to link the sample moments of the residuals to the empirical distribution of the residuals. This is then used to establish the properties of the cumulant based tests for normality referred to as the Jarque-Bera test.

The paper is organised so that the empirical process and the main result are presented in §2. Applications of this result to the above mentioned tests and plots are discussed in §3. The proofs are given in §4.

2 The empirical process

At first the sample average of residuals is discussed for first-order autoregressions with and without an intercept. This highlights the importance of the intercept when dealing with the empirical distribution of the residuals. Next, the general autoregressive model and the empirical distribution of the residuals are defined; finally, the asymptotic theory of the empirical distribution is given.

2.1 The role of the intercept

The intercept plays a very important role for the empirical distribution of autoregressive residuals. This point is best illustrated by examining the sample average of the residuals, which is the expectation of the empirical distribution.

Consider the first-order autoregression with an intercept as given by

$$X_t = \alpha X_{t-1} + \mu + \sigma \varepsilon_t \quad (t = 1, \dots, T), \quad (2.1)$$

conditional on X_0 and parameter space $\alpha, \mu, \sigma^2 \in \mathbf{R}^2 \times \mathbf{R}_+$. No assumptions are needed for the innovations, ε_t , at this point. Estimating the parameters by least squares regression leads to scaled residuals of the form

$$\widehat{\varepsilon}_t = \frac{X_t - \widehat{\alpha} X_{t-1} - \widehat{\mu}}{\widehat{\sigma}} \quad (t = 1, \dots, T).$$

A crucial property of the residuals, $\widehat{\varepsilon}_t$, is that their average is zero,

$$\frac{1}{T} \sum_{t=1}^T \widehat{\varepsilon}_t = 0, \quad (2.2)$$

due to the estimating equation for μ . This implies that the expectation of the empirical distribution of the residuals is zero. As an immediate consequence, the expectation of the empirical distribution is free of nuisance parameters. In line with this it will be shown in §2.3 that the empirical distribution is asymptotically

Gaussian and free of nuisance parameters. This has previously been shown for $|\alpha| < 1$ and $|\alpha| > 1$ by Pierce (1985) and by Koul and Leventhal (1989), respectively, whereas the result for $|\alpha| = 1$ has not been studied in the context of empirical processes.

In contrast, consider the first-order autoregression without intercept, satisfying

$$X_t = \alpha X_{t-1} + \sigma \varepsilon_t \quad (t = 1, \dots, T), \quad (2.3)$$

conditional on X_0 , with parameter space $\alpha, \sigma^2 \in \mathbf{R} \times \mathbf{R}_+$. The scaled least squares residuals are now of the form

$$\widehat{\varepsilon}_t = \frac{X_t - \widehat{\alpha} X_{t-1}}{\widehat{\sigma}} \quad (t = 1, \dots, T).$$

Since the intercept is excluded it will in general hold that $\sum_{t=1}^T \widehat{\varepsilon}_t \neq 0$. The application of standard asymptotic techniques, with details given in §4.1, reveals the following asymptotic behaviour for the sample average.

Theorem 2.1 *Suppose the model given by (2.3) is satisfied with independent innovations with distribution function F where $\mathbf{E}\varepsilon_t = 0$, $\mathbf{Var}\varepsilon_t = 1$, and $\mathbf{E}|\varepsilon_t|^{2+\gamma} < \infty$ for some $\gamma > 0$. Let $\mathbb{B}(u)$ be a standard Brownian motion. Then*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{\varepsilon}_t = \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_t + R,$$

where the remainder term R satisfies

$$\begin{aligned} R &\xrightarrow{\mathbf{P}} 0 && \text{for } \alpha \neq 1, \\ R &\xrightarrow{\mathbf{D}} \frac{\int_0^1 \mathbb{B}(u) d\mathbb{B}(u)}{\int_0^1 \{\mathbb{B}(u)\}^2 du} \int_0^1 \mathbb{B}(u) du && \text{for } \alpha = 1. \end{aligned}$$

This result indicates that in the absence of an intercept rather different asymptotic behaviour can be expected in the presence and absence of unit roots. Moving on to higher order autoregressions without an intercept, different Dickey-Fuller-type expressions arise depending on the number of unit roots. This wide range of limiting distributions is not encouraging. From a practical point of view the solution is simply to consider the studentised residuals instead of the scaled residuals, which of course amounts to the residuals of the regression (2.1). This is the solution adopted, implicitly, in the frequently used cumulant based tests for normality, see §3.4. For the remainder of the paper an intercept will therefore be included in the model. Note that this is a modeling assumption, not an assumption about the distribution. Asymptotic theory for empirical process for $|\alpha| < 1$, $\alpha = 1$ and $|\alpha| > 1$, have been found by Boldin (1981), Lee and Wei (1999) and Koul and Leventhal (1989), respectively. The result for $\alpha = -1$ has not been studied in the context of empirical processes.

2.2 The general autoregressive model

Let X_t be a p -dimensional time series partitioned in terms of a univariate time series Y_t and $(p - 1)$ -dimensional time series Z_t . The general univariate model is given by

$$Y_t = \rho Z_t + \sum_{j=1}^k \alpha_j Y_{t-j} + \sum_{j=1}^k \beta_j' Z_{t-j} + \nu D_{t-1} + \sigma \varepsilon_t, \quad (t = 1, \dots, T) \quad (2.4)$$

conditional on X_0, \dots, X_{1-k} , with independent innovations ε_t with distribution function F . The term D_{t-1} is a deterministic term, which will be discussed in further detail below. When ρ is restricted to zero, so Z_t is absent on the right hand side, this is the marginal equation of a vector autoregression, and when ρ is unrestricted this is an autoregressive distributed lags model. If the Z process is omitted this reduces to an autoregression. Note, that when $k = 0$ and $D_{t-1} = 1$ the model reduces to a classical regression model with an intercept. In addition, when ρ is restricted to zero, this becomes a location-scale model for Y_t .

Least squares estimation of the equation (2.4) gives the scaled residuals

$$\hat{\varepsilon}_t = \frac{Y_t - \hat{\rho} Z_t - \sum_{j=1}^k \hat{\alpha}_j Y_{t-j} - \sum_{j=1}^k \hat{\beta}_j' Z_{t-j} - \hat{\nu} D_{t-1}}{\hat{\sigma}},$$

if, for instance, ρ is unrestricted. The empirical distribution function of the residuals is defined as

$$\hat{F}(x) = \frac{1}{T} \sum_{t=1}^T 1_{(\hat{\varepsilon}_t \leq x)},$$

with the associated empirical process

$$\hat{\mathbb{F}}\{F(x)\} = \sqrt{T} \left\{ \hat{F}(x) - F(x) \right\}, \quad (2.5)$$

which has argument $u = F(x)$ on the unit interval.

In order to discuss the distribution of the empirical process the joint distribution of the time series $X_t = (Y_t, Z_t)'$ has to be specified. If this is assumed to satisfy a vector autoregression the results for general vector autoregressions given by Nielsen (2005) can be used. That paper is a generalisation of the work by Lai and Wei (1985), who did not consider deterministic terms. Thus, suppose the time series X_t and the deterministic series D_t satisfy the autoregressive equations

$$X_t = \sum_{j=1}^k A_j X_{t-j} + \mu D_{t-1} + \xi_t, \quad (t = 1, \dots, T), \quad (2.6)$$

$$D_t = \mathbf{D} D_{t-1}, \quad (2.7)$$

conditional on X_0, \dots, X_{1-k} , where the vector innovations, ξ_t , are partitioned as $\xi_t = (\xi_{y,t}, \xi_{z,t}')'$. The vector innovations are assumed to have mean zero and a positive definite variance matrix partitioned as

$$\text{Var}(\xi_t) = \Omega = \begin{pmatrix} \Omega_{yy} & \Omega_{yz} \\ \Omega_{zy} & \Omega_{zz} \end{pmatrix}. \quad (2.8)$$

The univariate model (2.4) can then arise in two ways:

1. When ρ is restricted to zero then (2.4) is simply the first equation in (2.6) so $\sigma\varepsilon_t = \xi_{y,t}$ with mean zero and variance $\sigma^2 = \Omega_{yy}$.
2. When $\rho = \Omega_{yz}\Omega_{zz}^{-1}$ and ξ is normally distributed, then (2.4) states the conditional model for Y_t given Z_t and the past, with $\sigma\varepsilon_t = \xi_{y,t} - \rho\xi_{z,t}$ with mean zero and variance $\sigma^2 = \Omega_{yy} - \Omega_{yz}\Omega_{zz}^{-1}\Omega_{zy}$.

The formulation for the deterministic term D_t allows a joint autoregressive companion representation of X_t, D_t , and is inspired by Johansen (2000). The matrix \mathbf{D} has characteristic roots on the complex unit circle, so D_t is a vector of terms such as a constant, a linear trend, or periodic functions like seasonal dummies. For example,

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix} \quad \text{with} \quad D_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

will generate a constant and a dummy for a bi-annual frequency. The deterministic term D_t is assumed to have linearly independent coordinates, which is formalised as follows.

Assumption 2.2 $|\text{eigen}(\mathbf{D})| = 1$ and $\text{rank}(D_1, \dots, D_{\dim \mathbf{D}}) = \dim \mathbf{D}$.

To ensure that an intercept is included in the model an additional assumption to \mathbf{D} is needed.

Assumption 2.3 Assume \mathbf{D} has at least one eigenvalue of unity.

2.3 Asymptotic theory for the empirical process

Before discussing the empirical process $\widehat{\mathbb{F}}$ itself, it is useful to consider the empirical process of the innovations, ε_t , for a moment. This is defined as

$$\widehat{\mathbb{U}}(u) = \frac{1}{\sqrt{T}} \sum_{t=1}^T [1_{\{\varepsilon_t \leq F^{-1}(u)\}} - u] = \frac{1}{\sqrt{T}} \sum_{t=1}^T [1_{\{F(\varepsilon_t) \leq u\}} - u], \quad (2.9)$$

for $0 \leq u \leq 1$. It is often referred to as the uniform empirical process since the variables $F(\varepsilon_t)$ are uniformly distributed. The asymptotic theory for the uniform process of the innovations is outlined in Billingsley (1968, §13) and is described in terms of a Brownian bridge. That is, under regularity conditions that will be introduced below,

$$\widehat{\mathbb{U}}(u) \xrightarrow{D} \mathbb{U}(u), \quad (2.10)$$

on the space $D[0, 1]$ of right continuous functions on the unit interval with limits from the left. Here \mathbb{U} is a standard Brownian Bridge, that is, a Gaussian process with expectation zero and covariance given by $\text{Cov}\{\mathbb{U}(u), \mathbb{U}(v)\} = u(1-v)$ for $u \leq v$. The Brownian bridge \mathbb{U} can be written in terms of a standard Brownian motion \mathbb{B} as $\mathbb{U}(u) = \mathbb{B}(u) - u\mathbb{B}(1)$. The asymptotic result for the empirical process $\widehat{\mathbb{F}}(u)$ is a generalisation of this result, where the limiting distribution is expressed in terms of stochastic integrals with respect to the Brownian bridge \mathbb{U} .

The properties of stochastic intergrals with respect to Brownian bridges are discussed by Shorack and Wellner (1986, p. 91-95). Stochastic integrals $\int_0^1 h_j(u) d\mathbb{U}(u)$, are well-defined for square integrable functions h_1, h_2 . They are normally distributed with expectation zero, and variance given by

$$\begin{aligned} & \text{Cov} \left\{ \int_0^1 h_1(u) d\mathbb{U}(u), \int_0^1 h_2(u) d\mathbb{U}(u) \right\} \\ &= \int_0^1 h_1(u) h_2(u) du - \left\{ \int_0^1 h_1(u) du \right\} \left\{ \int_0^1 h_2(u) du \right\}. \end{aligned} \quad (2.11)$$

The Brownian bridge itself can be written as $\mathbb{U}(v) = \int_0^1 1_{(u \leq v)} d\mathbb{U}(u)$.

To formulate the main result some assumptions to innovations ξ_t and ε_t are needed. The innovations, ξ_t , of the vector autoregression (2.6) are assumed to satisfy a martingale difference sequence assumption to exploit the consistency results of Nielsen (2005).

Assumption 2.4 *Suppose the sequence of innovations ξ_t of the vector autoregression (2.6) is a martingale difference sequence with respect to an increasing filtration \mathcal{F}_t so $\mathbb{E}(\xi_t | \mathcal{F}_{t-1}) = 0$ and where the initial values X_0, \dots, X_{1-k} are \mathcal{F}_0 -measurable and*

$$\sup_t \mathbb{E}\{(\xi_t' \xi_t)^{\lambda/2} | \mathcal{F}_{t-1}\} \stackrel{a.s.}{<} \infty \quad \text{for some } \lambda > 4, \quad (2.12)$$

$$\mathbb{E}(\xi_t \xi_t' | \mathcal{F}_{t-1}) \stackrel{a.s.}{=} \Omega \quad \text{where } \Omega \text{ is positive definite.} \quad (2.13)$$

A further set of three assumptions are needed for the innovations ε_t of the regression equation (2.4). It is not sufficient that these innovations are martingale difference sequences, essentially because the empirical process describes the entire distribution, and thus all moments of ε_t . The first assumption ensures the convergence of the uniform empirical process.

Assumption 2.5 *Suppose the innovations ε_t of the regression equation (2.4) are independent and identically distributed so $\mathbb{E}\varepsilon_t = 0$, $\text{Var}\varepsilon_t = 1$, and with marginal distribution function F .*

A second assumption is a martingale assumption, which is needed for the case where ρ is unrestricted so the regression equation (2.4) becomes an autoregressive distributed lags equation.

Assumption 2.6 *If ρ is unrestricted define the filtration \mathcal{G}_{t-1} as the sigma field over \mathcal{F}_{t-1} and Z_t , otherwise, if ρ is restricted to zero let $\mathcal{G}_{t-1} = \mathcal{F}_{t-1}$. Suppose the innovations ε_t are independent of \mathcal{G}_{t-1} .*

The third assumption concerns the distribution function F of the innovations, ε_t .

Assumption 2.7 *Suppose the distribution function F has density f which is positive and differentiable everywhere, and satisfies*

$$\sup_{x \in \mathbf{R}} |xf(x)| < \infty, \quad \sup_{x \in \mathbf{R}} |f'(x)| < \infty, \quad \sup_{x \in \mathbf{R}} |x^2 f'(x)| < \infty.$$

The Assumption 2.7 is satisfied by many distributions, notably the standard normal distributions. Denoting the standard normal density by φ it holds that $\varphi'(x) = -x\varphi(x)$. Since φ has exponentially declining tails the boundedness follows.

The main result can now be formulated. This shows that as long as an intercept is included in the autoregression the same Gaussian limit distribution applies for all values of the characteristic roots. Thus, it is possible to make inference about the distribution of the autoregressive innovations without knowing the location of the characteristic roots.

Theorem 2.8 *Suppose the model (2.6) and the Assumptions 2.2, 2.3, 2.4, 2.5, 2.6, 2.7 are satisfied. Then*

$$\widehat{\mathbb{F}} \{F^{-1}(\cdot)\} \xrightarrow{D} \mathbb{X}_F(\cdot),$$

on $D[0, 1]$, where \mathbb{X}_F is the Gaussian process defined by

$$\mathbb{X}_F(u) = \left[\begin{array}{c} 1 \\ f \{F^{-1}(u)\} \\ \frac{1}{2} F^{-1}(u) f \{F^{-1}(u)\} \end{array} \right]' \int_0^1 \left[\begin{array}{c} 1_{[s \leq u]} \\ F^{-1}(s) \\ \{F^{-1}(s)\}^2 \end{array} \right] d\mathbb{U}(s). \quad (2.14)$$

The process \mathbb{X}_F defined in (2.14) is the sum of three components. The first component is simply the Brownian Bridge $\mathbb{U}(u)$ arising from the uniform empirical process $\widehat{\mathbb{U}}$. The second and the third components relate to $\widehat{\mathbb{F}} - \widehat{\mathbb{U}}$ and arise due to the estimation of the variance and the intercept, respectively. The process \mathbb{X}_F is standard, in the sense of applying to cross-sectional regression problems including a constant, see Shorack & Wellner (1986, p.197f).

The process \mathbb{X}_F can be simulated in various ways. The first approach is to use the covariance structure of \mathbb{X}_F . For any choice of reference distribution F this covariance structure can be found using the formula (2.11). For the Gaussian case it is found in Theorem 2.10 below. For a particular grid over the interval $[0, 1]$, the covariance matrix is computed and its square root multiplied onto a vector of generated independent standard normal variables. The second approach is to compute the stochastic integrals in (2.14) directly. First, a Brownian motion \mathbb{B} is computed for a grid over $[0, 1]$ by taking partial sums of generated independent standard normal variables and is transformed into a Brownian bridge using the formula $\mathbb{U}(u) = \mathbb{B}(u) - u\mathbb{B}(1)$. Next, the integral (2.14) is formed. This approach is a little more convoluted, but numerically faster when dealing with a fine grid over $[0, 1]$.

2.4 Empirical moments

The asymptotic properties of the moments of the residuals can be found rather easily from the empirical process. This point seems to have escaped the literature. It will later, in §3.4, be used to establish the asymptotic behaviour of the usual cumulant-based tests for normality. Some notation is needed. Define the empirical moments of the standardised residuals,

$$\hat{\mu}_m = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t^m = \int_{\mathbf{R}} x^m d\hat{\mathbf{F}}(x),$$

and the least squares estimator for the variance parameter, normalised by T ,

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \left(Y_t - \hat{\rho}Z_t - \sum_{j=1}^k \hat{\alpha}_j Y_{t-j} - \sum_{j=1}^k \hat{\beta}'_j Z_{t-j} - \hat{\nu}D_{t-1} \right)^2.$$

It then holds that $\hat{\mu}_2 = 1$, and, when including an intercept in the model, also $\hat{\mu}_1 = 0$. Defining the population moments

$$\mu_m = \int_{\mathbf{R}} x^m d\mathbf{F}(x) = \int_0^1 \{\mathbf{F}^{-1}(u)\}^m du, \quad (2.15)$$

the sample and population moments can be brought together as

$$\begin{aligned} \sqrt{T}(\hat{\mu}_m - \mu_m) &= \sqrt{T} \left\{ \int_{\mathbf{R}} x^m d\hat{\mathbf{F}}(x) - \int_{\mathbf{R}} x^m d\mathbf{F}(x) \right\} \\ &= \int_{\mathbf{R}} x^m d\hat{\mathbb{F}}_T(x) = \int_0^1 \{\mathbf{F}^{-1}(u)\}^m d\hat{\mathbb{F}}_T\{\mathbf{F}^{-1}(u)\}. \end{aligned} \quad (2.16)$$

The asymptotic theory then follows by replacing the empirical process $\hat{\mathbb{F}}_T$ with the limiting Gaussian process $\mathbb{X}_{\mathbf{F}}$ established in Theorem 2.8. To describe the resulting limiting normal distribution let

$$\begin{aligned} h_2(u) &= \{\mathbf{F}^{-1}(u)\}^2, \\ h_m(u) &= \{\mathbf{F}^{-1}(u)\}^m - m\mu_{m-1}\{\mathbf{F}^{-1}(u)\} - \frac{m}{2}\mu_m\{\mathbf{F}^{-1}(u)\}^2. \end{aligned}$$

Theorem 2.9 *Suppose the model (2.6) and the Assumptions 2.2, 2.3, 2.4, 2.5, 2.6, 2.7 are satisfied, and that \mathbf{F} has moments of order at least $2M$. Then*

$$\sqrt{T}(\hat{\sigma}^2 - \sigma^2, \hat{\mu}_3 - \mu_3, \dots, \hat{\mu}_M - \mu_M)$$

converge jointly in distribution to a normal distribution with mean zero. The limiting distribution can be represented as

$$\sqrt{T} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1 \right) \stackrel{\mathbb{D}}{\rightarrow} \int_0^1 h_2(u) d\mathbb{U}(u) = \int_0^1 \{\mathbf{F}^{-1}(u)\}^2 d\mathbb{U}(u), \quad (2.17)$$

$$\sqrt{T}(\hat{\mu}_m - \mu_m) \stackrel{\mathbb{D}}{\rightarrow} \int_0^1 h_m(u) d\mathbb{U}(u) = \int_0^1 \{\mathbf{F}^{-1}(u)\}^m d\mathbb{X}_{\mathbf{F}}(u), \quad (2.18)$$

and the covariance matrix has entries

$$\omega_{mn} = \text{Cov} \left\{ \int_0^1 h_m(u) d\mathbb{U}(u), \int_0^1 h_n(u) d\mathbb{U}(u) \right\}$$

which satisfy, for $m, n \geq 3$,

$$\begin{aligned} \omega_{22} &= \mu_4 - 1, & \omega_{2m} &= \mu_{m+2} - m\mu_3\mu_{m-1} - \mu_m \left\{ 1 + \frac{m}{2}(\mu_4 - 1) \right\}, \\ \omega_{mn} &= \mu_{m+n} - \mu_{m+1}\mu_{n+1} - \left(\frac{mn}{2} + 1 \right) \mu_m\mu_n + \frac{mn}{4}(\mu_4 - 3)\mu_m\mu_n \\ &\quad + (\mu_{m+1} - m\mu_{m-1})(\mu_{n+1} - n\mu_{n-1}) + \frac{mn}{2}\mu_3(\mu_{m-1}\mu_n + \mu_m\mu_{n-1}) \\ &\quad - \frac{n}{2}\mu_n \{ \mu_{m+2} - (m+1)\mu_m \} - \frac{m}{2}\mu_m \{ \mu_{n+2} - (n+1)\mu_n \}. \end{aligned}$$

The result can form the basis for a cumulant-based test of the reference distribution F in a location-scale model. For instance, if ε_t follows a standard Gaussian distribution, a Laplace distribution, or a t -distribution with known degrees of freedom, then $Y_t = \mu + \sigma\varepsilon_t$ is a location-scale model. The Gaussian case is studied in further detail in §2.5.

2.5 The Gaussian case

The case where the reference distribution F is standard Gaussian, denoted Φ is of special interest. At first, the covariance structure of the limiting Gaussian process \mathbb{X}_F is found. Subsequently, the moments of the empirical distribution are described.

Theorem 2.10 *Suppose $F(u) = \Phi(u)$ is the standard normal distribution function with density φ . Then $\mathbb{X}_F = \mathbb{X}_\Phi$ has covariance given by*

$$\text{Cov} \{ \mathbb{X}_\Phi(u), \mathbb{X}_\Phi(v) \} = u(1-v) - \varphi \{ \Phi^{-1}(u) \} \varphi \{ \Phi^{-1}(v) \} \left\{ \frac{\Phi^{-1}(u)\Phi^{-1}(v)}{2} + 1 \right\},$$

for $0 \leq u \leq v \leq 1$, or equivalently, for $x, y \in \mathbf{R}$ so $x \leq y$,

$$\text{Cov} [\mathbb{X}_\Phi \{ \Phi(x) \}, \mathbb{X}_\Phi \{ \Phi(y) \}] = \Phi(x) \{ 1 - \Phi(y) \} - \varphi(x) \varphi(y) \left(\frac{xy}{2} + 1 \right).$$

Theorem 2.9 presented the asymptotic distribution of the sample moments of the standardised residuals. This result simplifies considerably in the Gaussian case. In that case the population moments are

$$\mu_m = \int_{\mathbf{R}} x^m d\Phi(x) = \int_0^1 \{ \Phi^{-1}(u) \}^m du = \begin{cases} 0 & \text{for } m \text{ odd,} \\ (m-1)!! & \text{for } m \text{ even,} \end{cases} \quad (2.19)$$

where $(m-1)!! = (m-1)(m-3)\cdots 3 \cdot 1$ is the odd factorial. In particular $\mu_3 = 0$, $\mu_4 = 3$, and $\mu_{m+1} = m\mu_{m-1}$. This immediately gives the following result, noting that Assumptions 2.5 and 2.7 are satisfied in the normal case.

Corollary 2.11 *Suppose the model (2.6) and the Assumptions 2.2, 2.3, 2.4, 2.6 are satisfied and the innovations ε_t are standard normal so $\mathbf{F} = \Phi$. Then, for any M*

$$\sqrt{T} \left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1, \widehat{\mu}_3 - \mu_3, \dots, \widehat{\mu}_M - \mu_M \right)$$

converge jointly in distribution to a normal distribution with mean zero. The covariance matrix of the limiting distribution has entries, for $m, n \geq 3$,

$$\omega_{22} = 2, \quad \omega_{2m} = 0, \quad \omega_{mn} = \mu_{m+n} - \mu_{m+1}\mu_{n+1} - \left(\frac{mn}{2} + 1 \right) \mu_m \mu_n.$$

The property that $\omega_{2m} = 0$ shows that $\widehat{\sigma}^2$ is asymptotically independent of all higher moments of the standardised residuals, which in turn implies that it is independent of the higher cumulants. This relates to the exact independence found by Fisher (1930) for samples of independent normal variates. Note also that $\omega_{mn} = 0$ when m is even and n is odd, implying that for instance the third and fourth cumulants of the residuals are asymptotically independent.

3 Applications

Theorem 2.8 has a wide range of applications that can be helpful in autoregressive analysis. In the following, the empirical process is at first transformed into a standardised empirical process and a quantile process. Next, Kolmogorov-Smirnov and Anderson-Darling tests are discussed. Then confidence bands are presented for Probability-Probability and Quantile-Quantile plots. These are frequently used in econometrics, but without confidence bands. Finally, the frequently used cumulant based tests for normality are discussed.

3.1 The standardised empirical process and the quantile process

In Theorem 2.8, a Gaussian approximation was found for the distribution of the empirical process. Using this result, a standardised version of the empirical process can be defined

$$\widehat{\mathbb{Z}}_{\mathbf{F}}(u) = \frac{\widehat{\mathbb{F}}(u)}{\widehat{\text{std}}\{\mathbb{X}_{\mathbf{F}}(u)\}} \quad \text{for } u \in (0, 1).$$

This process is not defined at the end points $u = 0$ and $u = 1$ since both the empirical process and its variance are then zero. For each u in the open interval $(0, 1)$ Theorem 2.8 implies the pointwise convergence result:

$$\widehat{\mathbb{Z}}_{\mathbf{F}}(u) \xrightarrow{\text{D}} \mathbb{Z}_{\mathbf{F}}(u) = \frac{\mathbb{X}_{\mathbf{F}}(u)}{\text{std}\{\mathbb{X}_{\mathbf{F}}(u)\}}. \quad (3.1)$$

The sequence of standardised empirical processes, $\widehat{\mathbb{Z}}_{\mathbf{F}}$, is, however, not tight. The issue is the variation near 0 and near 1. In this way, Čibisov (1966) showed that

$$\sup_{0 < u < 1} \left| \frac{\widehat{\mathbb{U}}(u)}{\text{std}\{\mathbb{U}(u)\}} \right| \xrightarrow{\text{P}} \infty.$$

The rate of divergence has been studied intensively, with an overview given in Shorack and Wellner (1986). The process \widehat{Z}_F will, however, converge on $D[b, c]$ where $[b, c]$ is an arbitrary closed subinterval of $(0, 1)$.

Rather than looking at the distribution function for the residuals, it is often of interest to look at their quantiles, defined in terms of the inverse distribution function, \widehat{F}^{-1} . The standardised quantile process is defined as

$$\widehat{Q}_F(x) = f(x) \sqrt{T} \left[\widehat{F}^{-1} \{F(x)\} - x \right]. \quad (3.2)$$

Using the functional δ -method, in an argument made more formally by Shorack and Wellner (1986, Chapter 18), it holds that

$$\widehat{Q}_F(x) \xrightarrow{D} Q_F(x), \quad (3.3)$$

for each point $x \in \mathbf{R}$, as well as $\widehat{Q}_F\{F^{-1}(\cdot)\}$ converges on $D[b, c]$ where $[b, c] \subset (0, 1)$. The limiting process has the same covariance process as \mathbb{X}_F due to the normalisation with f in (3.2). Dividing through by $f(x)$ gives the pointwise result

$$\widehat{R}_F(x) = \sqrt{T} \left[\widehat{F}^{-1} \{F(x)\} - x \right] \xrightarrow{D} \frac{Q_F(x)}{f(x)}. \quad (3.4)$$

3.2 Test statistics based on the empirical process

In many applications it is convenient to summarize the empirical process in a single statistic. There are two reasons for this. *First*, for the uniform empirical process, distribution free tests can be constructed based on, for instance, the Kolmogorov-Smirnov statistic. This does, however, not seem possible when the location and scale are unknown as here. *Secondly*, tests can be constructed by evaluating the empirical process at a single point. Looking at two different points, x_1 and x_2 , the test statistics will be dependent. To achieve a correct size this has to be taken into account. This is done by looking at summary statistics such as the Kolmogorov-Smirnov statistic. Different summary statistics are found depending on the choice of weight function that is applied to the empirical process over the points $x \in \mathbf{R}$. There is no optimal way of choosing a weight function. Likewise, such statistics are bound to be dependent, so when using several statistics it is not clear how to achieve a correct size.

Kolmogorov-Smirnov-type statistics can be formed from the empirical process by

$$\widehat{D}_F = \sup_{x \in \mathbf{R}} \left| \widehat{\mathbb{F}} \{F(x)\} \right| = \sup_{0 \leq u \leq 1} \left| \widehat{\mathbb{F}}(u) \right| \xrightarrow{D} \sup_{0 \leq u \leq 1} |\mathbb{X}_F(u)| = D_F, \quad (3.5)$$

$$\widehat{D}_F^+ = \sup_{x \in \mathbf{R}} \widehat{\mathbb{F}} \{F(x)\} = \sup_{0 \leq u \leq 1} \widehat{\mathbb{F}}(u) \xrightarrow{D} \sup_{0 \leq u \leq 1} \mathbb{X}_F(u) = D_F^+, \quad (3.6)$$

where the limiting distributions arise by applying the Continuous Mapping Theorem to the asymptotic distribution for the empirical process reported in Theorem 2.8. These statistics put most weight on values of u closest to 0.5. The limiting

	E	Var	E log	Var log	50%	80%	90%	95%	97.5%	99%
D_{Φ}	0.631	0.0219	-0.487	0.0526	0.612	0.748	0.830	0.903	0.971	1.053
D_{Φ}^+	0.547	0.0231	-0.640	0.0731	0.524	0.666	0.753	0.831	0.903	0.990

Table 1: Simulated distribution of \widehat{D}_{Φ} and \widehat{D}_{Φ}^+

distributions of D_F , D_F^+ depend on the distribution of F . This is in contrast to the situation where the empirical distribution of independent, identically variables is compared to a known distribution. In that situation only the uniform term \widehat{U} is of relevance and a time transformation argument can be employed.

When testing for normality the distributions D_{Φ} and D_{Φ}^+ are of relevance. These were previously reported by Stephens (1974, Table 1A, case 3). Stephens' experiment was repeated using a fine grid with 10^4 points and 10^6 repetitions giving the numbers reported in Table 1. A convenient approximation to the p-values can be found using a Gamma distribution with the reported mean and variance as done for instance for a Dickey-Fuller F-type distribution in Nielsen (1997). Due to the extreme value nature of the Kolmogorov-Smirnov statistic, the Gamma approximation will under-fit the extreme upper tail slightly. For the reported quantiles the relative error of the Gamma quantiles compared to the simulated quantiles was at most 3.0%, and at most 1.8% when excluding the two most extreme quantiles. It was also attempted to fit a Weibull distribution, as the asymptotic distribution of the D_F^+ statistic based on the uniform empirical process is Weibull, see Billingsley (1968, p.85). The Weibull distribution can be fitted using mean and variance of the log-transformed statistic, see Johnson, Kotz, and Balakrishnan (1994, §21.4). For the empirical process of residuals the fit is, however, much worse than the Gamma fit, with relative errors of up to 17%.

While the Kolmogorov-Smirnov statistic puts most weight on deviations in the middle of the distribution, Anderson and Darling (1952) considered the possibility of constructing Kolmogorov-Smirnov-type statistics for the standardised empirical process, \widehat{Z}_F , see (3.1). Anticipating the result of Čibisov (1966), they suggested taking supremum over a closed interval in the interior of the unit interval. Considering a symmetric interval for simplicity, this gives

$$\widehat{K}_{F,a} = \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} \left| \widehat{Z}_F(u) \right| \xrightarrow{D} \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} |Z_F(u)| = K_{F,a}.$$

A Kolmogorov-Smirnov-type statistics can be constructed in the same way for the empirical quantile process

$$\begin{aligned} \widehat{Q}_{F,a} &= \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} \left| \frac{\widehat{Q}_F(u)}{\widehat{\text{std}} \left\{ \widehat{Q}_F(u) \right\}} \right| = \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} \left| \frac{\widehat{R}_F(u)}{\widehat{\text{std}} \left\{ \widehat{R}_F(u) \right\}} \right| \\ &\xrightarrow{D} \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} |Z_F(u)| = K_{F,a}, \\ \widehat{R}_{F,a} &= \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} \left| \widehat{R}_F(u) \right| \xrightarrow{D} \sup_{\frac{1}{2}-a < u < \frac{1}{2}+a} |R_F(u)| = R_{F,a}. \end{aligned}$$

		1	a	a^2	a^{-1}	a^{-2}	$(\frac{1}{2} - a)^{-1}$	$(\frac{1}{2} - a)^{-2}$
$K_{\Phi,a}$	E	1.969	1.405		-0.04406	0.0008920	0.00028	
	Var	0.365	-0.291	0.141			-0.00001	
$R_{\Phi,a}$	E	1.672	-1.583	4.904	-0.02773		0.03722	-0.0000677
	Var	0.135	0.053	-1.293			0.02743	-0.0000151

Table 2: Response surface in a for expectation and variance of $\widehat{K}_{\Phi,a}$ and $\widehat{Q}_{\Phi,a}$.

When testing for normality the distributions $K_{\Phi,a}$ and $R_{\Phi,a}$ are of relevance. These were simulated for a range of a values, a fine grid over u with 10^4 points and 10^6 repetitions. Response surfaces in a for the expectation and variance of $K_{\Phi,a}$ and $R_{\Phi,a}$ are reported in Table 2 using 22 values of a chosen as (0.05:0.40, 0.05), (0.41:0.49, 0.01), (0.491:0.495, 0.001). In all cases, the R^2 of the fits exceed 0.9995. It is not advisable to extrapolate the response surface for values of a outside $0.01 < a < 0.495$. For a given value of a the expectation and variance are computed and the distribution approximated using a Gamma distribution as above. It is evident that the response surfaces diverge for $a \rightarrow 0.5$. Different choices of a will emphasize different departures from normality. Due to the Čibisov result the literature does not give any guidance towards the choice of a as a function of the sample size.

3.3 P-P and Q-Q plots

At present, probability-probability and quantile-quantile plots, also called P-P and Q-Q plots, are used without confidence bands. Based on the work present here, two types of confidence bands can be derived: pointwise bands and simultaneous bands based on the statistics presented in §3.2. Pointwise bands would be used in situations where the nature of the departure from the reference distribution is unknown, whereas the simultaneous bands are used to detect more specific types of departures.

Probability-probability plots are plots of $\widehat{u} = \widehat{F}(x)$ against $u = F(x)$ on a $[0, 1] \times [0, 1]$ -square. Confidence bands of the type $u \pm n^{-1/2} c_\alpha \sigma_u$ can be constructed using the results from above. Three different bands are worth noting.

Pointwise confidence bands can be established from Theorem 2.8, where c_α is the $1 - \alpha/2$ quantile of the standard normal distribution and σ_u^2 is the variance $\text{Var}\{\mathbb{X}_F(u)\}$ at the point u . For the standard normal case, $F = \Phi$, the variance is reported in Theorem 2.10. For other reference distributions it can be computed using (2.11).

Kolmogorov-Smirnov bands can be constructed by $\sigma_u = 1$ and choosing c_α from limiting distributions of the Kolmogorov-Smirnov statistics. This gives bands that are straight lines parallel to the 45° -line. This emphasises departures in the middle of the distribution. Bands of “two-sided” nature are found by choosing c_α as the $1 - \alpha$ quantile of D_F , whereas a “one-sided” are found from the $1 - \alpha/2$ quantile of D_F^+ .

Anderson-Darling bands are constructed by first choosing a value of a (perhaps

0.3, 0.4, or 0.45). Then σ_u is chosen as for the pointwise bands, whereas c_α is chosen as the $1 - \alpha$ quantile of $K_{\Phi, a}$.

Quantile-quantile plots are plots of $\hat{x} = \hat{F}^{-1}(u)$ against $x = F^{-1}(u)$ on a $\mathbf{R} \times \mathbf{R}$ -square. Confidence bands of the type $u \pm n^{-1/2}c_\alpha\sigma_u$ can be constructed. Two different bands are worth noting.

Pointwise confidence bands can be established directly from Theorem 2.8 and (3.4). Here c_α is the $1 - \alpha/2$ quantile of the standard normal distribution and σ_u^2 equals $\text{Var}[\mathbb{X}_F\{F(x)\}]/\{f(x)\}^2$ at the point x . For the standard normal case, $F = \Phi$, the variance is reported in Theorem 2.10.

Kolmogorov-Smirnov bands can be constructed by $\sigma_u = 1$ and choosing c_α from the $1 - \alpha$ quantile of R_F .

3.4 Test for normality based on cumulants

Assuming an intercept is included in the model, the empirical cumulants can be computed from the sample moments of the standardised residuals, so

$$\hat{\kappa}_3 = \hat{\mu}_3, \quad \hat{\kappa}_4 = \hat{\mu}_4 - 3.$$

The following result follows directly from Corollary 2.11.

Corollary 3.1 *Suppose the model (2.6) and the Assumptions 2.2, 2.3, 2.4, 2.6 are satisfied, and that the innovations ε_t are $\mathbf{N}[0, 1]$ -distributed so $F(u) = \Phi(u)$. Then*

$$\left(\hat{\kappa}_3\sqrt{T/6}, \hat{\kappa}_4\sqrt{T/24}\right) \xrightarrow{D} \mathbf{N}_2[0, I_2].$$

Note that Assumption 2.3 is trivially satisfied, since the cumulants are based on the central moments, regardless of whether an intercept is included in the model. The Theorem presented here requires that the innovations ε_t are independent as stated in Assumption 2.5. This assumption is needed to be able to work with the empirical process, but could be relaxed with a proof based on direct expansions of the cumulants.

For models with independent observations, such a result was first established by Thiele (1889), see Lauritzen (2002) for a translation, and later by Pearson (1902), who argued that $(T/6)^{1/2}\hat{\kappa}_3$ and $(T/24)^{1/2}\hat{\kappa}_4$ are asymptotically standard normal. A large literature has investigated the joint behaviour in finite samples of independent observations. Jarque and Bera (1987) have given these tests a likelihood based motivation. Recently Kilian and Demiroglu (2002) have proved that these results hold for a class of cointegrated vector autoregressions with known cointegration rank. The proofs of these results are based on direct expansions of the sample moments of $\hat{\sigma}\hat{\varepsilon}_t$ rather than appealing to the empirical process of $\hat{\varepsilon}_t$. While the argument based on the empirical process requires independent innovations, an argument based on expansions of the sample moments could be made for innovations arising from a martingale difference sequence with constant conditional moments up to a certain order.

For cointegration applications the result of Kilian and Demiroglu (2002) can be combined with the present result in an interesting way. In a first step an

unrestricted vector autoregression is fitted to the data and the hypothesis of normality is tested using the new result, where no knowledge of the autoregressive parameters is required. Secondly, a cointegration analysis is performed using the likelihood procedure of Johansen (1995). Once the cointegration rank has been determined and imposed, the hypothesis of normality can then be tested once again for the restricted model using the results of Kilian and Demiroglu (2002).

4 Proofs

4.1 The sample average

Proof of Theorem 2.1. First, $\widehat{\sigma}\widehat{\varepsilon}_t = \sigma\varepsilon - (\widehat{\alpha} - \alpha)X_{t-1}$ by the model equation (2.6), so

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{\varepsilon}_t = \frac{\sigma}{\widehat{\sigma}} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_t - \widetilde{R} \right) \quad \text{where} \quad \widetilde{R} = \frac{(\widehat{\alpha} - \alpha)}{\sigma\sqrt{T}} \sum_{t=1}^T X_{t-1}.$$

Now, \widetilde{R} equals the product $\widetilde{R}_1\widetilde{R}_2$ where

$$\widetilde{R}_1 = \frac{\sum_{t=1}^T X_{t-1}\varepsilon_t}{\sqrt{\sum_{t=1}^T X_{t-1}^2}}, \quad \widetilde{R}_2 = \frac{\sum_{t=1}^T X_{t-1}}{\sqrt{T \sum_{t=1}^T X_{t-1}^2}}.$$

The estimator $\widehat{\sigma}$ is consistent, whereas $\widetilde{R}_1 = o(T^{1/2-\eta})$ *a.s.* for some $\eta > 0$ for all α (Nielsen, 2005, §2). The result for $\alpha \neq 1$ follows by showing that $\widetilde{R}_2 = o(T^{\eta-1/2})$ *a.s.*

When $|\alpha| < 1$ then $\widetilde{R}_2 = o(T^{\eta-1/2})$ *a.s.* (Nielsen, 2005, Theorem 6.4).

When $|\alpha| > 1$ then $\widetilde{R}_2 = O(T^{-1/2})$ *a.s.* (Nielsen, 2005, Theorem 9.1).

When $\alpha = -1$ then $\widetilde{R}_2 = o(T^{\eta-1/2})$ *a.s.* To see this, write X_{t-1} as the sum of $(-1)^{t-1}X_0$ and $\sigma \sum_{s=1}^{t-1} (-1)^{t-1-s} \varepsilon_s$, so $\sum_{t=1}^T X_{t-1}$ is the sum of $X_0 \sum_{t=0}^{T-1} (-1)^t = X_0\{1 - (-1)^T\}/2$ and $\sigma \sum_{s=1}^{T-1} \varepsilon_s \sum_{t=0}^{T-s-1} (-1)^t = \sigma \sum_{s=1}^{T-1} \varepsilon_s \{1 - (-1)^{T-s}\}/2$. Thus, $\sum_{t=1}^T X_{t-1} = o(T^{1/2+\eta})$ *a.s.* by the Law of Iterated Logarithms. As for the denominator, $(T \sum_{t=1}^T X_{t-1}^2)^{-1/2} = O(T^{-1})$ *a.s.*, see Nielsen (2005, Theorem 8.1).

Finally, when $\alpha = 1$ write $\sigma\widetilde{R}$ as the product of $T^{-1} \sum_{t=1}^T X_{t-1}\varepsilon_t$, $T^{-3/2} \sum_{t=1}^T X_{t-1}$, and $(T^{-2} \sum_{t=1}^T X_{t-1}^2)^{-1}$ and apply standard unit root asymptotics. ■

4.2 Proof of main theorem

Theorem 2.8 is now proved. For the asymptotic analysis, it is convenient to decompose the empirical process. To facilitate this, the set $(\widehat{\varepsilon}_t \leq x)$ is rewritten in three steps: First, both sides of the inequality are scaled by $\widehat{\sigma}/\sigma$ to bring the residuals to the population scale; secondly, $\varepsilon_t - \widehat{\sigma}\widehat{\varepsilon}_t/\sigma$ is added to both sides; and thirdly, x is added and subtracted on the right. This gives

$$(\widehat{\varepsilon}_t \leq x) = (\varepsilon_t \leq x + \widehat{z}_t), \quad (4.1)$$

where

$$\widehat{z}_t = \widehat{a}x + \widehat{b}_t, \quad \text{where} \quad \widehat{a} = \frac{\widehat{\sigma}}{\sigma} - 1, \quad \widehat{b}_t = \varepsilon_t - \frac{\widehat{\sigma}}{\sigma}\widehat{\varepsilon}_t. \quad (4.2)$$

The empirical distribution of the residuals can then be decomposed as

$$\widehat{\mathbb{F}}(u) = \widehat{\mathbb{U}}(u) + \widehat{\mathbb{V}}(u) + \widehat{\mathbb{W}}(u), \quad (4.3)$$

where

$$\begin{aligned} \widehat{\mathbb{U}}(u) &= \frac{1}{\sqrt{T}} \sum_{t=1}^T [1_{\{\varepsilon_t \leq F^{-1}(u)\}} - u], \\ \widehat{\mathbb{V}}(u) &= f\{F^{-1}(u)\} \frac{1}{\sqrt{T}} \sum_{t=1}^T \widehat{z}_t, \\ \widehat{\mathbb{W}}(u) &= \frac{1}{\sqrt{T}} \sum_{t=1}^T [1_{\{\widehat{\varepsilon}_t \leq F^{-1}(u)\}} - 1_{\{\varepsilon_t \leq F^{-1}(u)\}} - f\{F^{-1}(u)\} \widehat{z}_t]. \end{aligned}$$

The components $\widehat{\mathbb{U}}$ and $\widehat{\mathbb{V}}$ are the leading terms. They are analysed in Theorem 4.2 below. For the analysis of $\widehat{\mathbb{V}}$ it is convenient to decompose $\widehat{\mathbb{V}} = \widehat{\mathbb{V}}_1 + \widehat{\mathbb{V}}_2$, using the definition of \widehat{z}_t in (4.1). First, let

$$\widehat{\mathbb{V}}_1(u) = f\{F^{-1}(u)\} \frac{1}{\sqrt{T}} \sum_{t=1}^T \left(\varepsilon_t - \frac{\widehat{\sigma}}{\sigma}\widehat{\varepsilon}_t \right) = f\{F^{-1}(u)\} \frac{1}{\sqrt{T}} \sum_{t=1}^T \varepsilon_t, \quad (4.4)$$

since $\sum_{t=1}^T \widehat{\varepsilon}_t = 0$ when an intercept is included in the model, as stipulated in Assumption 2.3, whereas, letting $F(x) = u$,

$$\widehat{\mathbb{V}}_2(u) = F^{-1}(u) f\{F^{-1}(u)\} \sqrt{T} \left(\frac{\widehat{\sigma}}{\sigma} - 1 \right). \quad (4.5)$$

The third component, $\widehat{\mathbb{W}}$, vanishes. For the analysis of $\widehat{\mathbb{W}}$ it is convenient to decompose $\widehat{\mathbb{W}} = \widehat{\mathbb{W}}_1 + \widehat{\mathbb{W}}_2 + \widehat{\mathbb{W}}_3$, letting $F(x) = u$,

$$\begin{aligned} \widehat{\mathbb{W}}_1(u) &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \{1_{(\varepsilon_t \leq x + \widehat{a}x)} - 1_{(\varepsilon_t \leq x)} - f(x)\widehat{a}x\}, \\ \widehat{\mathbb{W}}_2(u) &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \{F(x + \widehat{a}x + \widehat{b}_t) - F(x + \widehat{a}x) - f(x)\widehat{b}_t\} \\ \widehat{\mathbb{W}}_3(u) &= \frac{1}{\sqrt{T}} \sum_{t=1}^T \{1_{(\varepsilon_t \leq x + \widehat{a}x + \widehat{b}_t)} - 1_{(\varepsilon_t \leq x + \widehat{a}x)} - F(x + \widehat{a}x + \widehat{b}_t) + F(\widehat{a}x)\}. \end{aligned}$$

The first term deals with the estimation of the scale parameter σ , while the other two deal with the estimation of expectation parameters. The first and the third term are referred to as uniform asymptotic linearity properties by Koul (2002), whereas the second is a Taylor expansion. It is shown in Theorems 4.3, 4.5, 4.6

that these terms disappear. The proof of Theorem 2.8 can now be summarised as follows.

Proof of Theorem 2.8. Decompose $\widehat{\mathbb{F}} = \widehat{\mathbb{U}} + \widehat{\mathbb{V}}_1 + \widehat{\mathbb{V}}_2 + \widehat{\mathbb{W}}_1 + \widehat{\mathbb{W}}_2 + \widehat{\mathbb{W}}_3$. Theorem 4.2 shows the convergence of the leading terms $\widehat{\mathbb{U}}, \widehat{\mathbb{V}}_1, \widehat{\mathbb{V}}_2$. Theorems 4.3, 4.5, 4.6 show that $\widehat{\mathbb{W}}_1, \widehat{\mathbb{W}}_2, \widehat{\mathbb{W}}_3$ vanish. ■

Before looking at the asymptotic behaviour of $\widehat{\mathbb{U}}$ and $\widehat{\mathbb{V}}$ it is convenient to consider the estimators for the covariance parameters.

Lemma 4.1 *It holds that*

$$\frac{\widehat{\sigma}}{\sigma} - 1 = \frac{1}{2} \left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1 \right) + \mathcal{O} \left\{ \left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1 \right)^2 \right\}.$$

Suppose model (2.6) and Assumptions 2.2, 2.4 are satisfied. Then

$$\sqrt{T} \left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1 \right) = \frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_t^2 - 1) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1), \quad (4.6)$$

as well as $\sqrt{T}(\widehat{\rho} - \rho) = \mathcal{O}_{\mathbb{P}}(1)$.

Proof of Lemma 4.1. The first result follows from the Taylor expansion

$$\frac{\widehat{\sigma}}{\sigma} - 1 = \sqrt{1 + \frac{\widehat{\sigma}^2 - \sigma^2}{\sigma^2}} - 1 = \frac{\widehat{\sigma}^2 - \sigma^2}{2\sigma^2} + \mathcal{O}_{\mathbb{P}} \left\{ \left(\frac{\widehat{\sigma}^2 - \sigma^2}{\sigma^2} \right)^2 \right\}.$$

Next, Nielsen (2005, Theorem 2.6) shows that under Assumptions 2.2, 2.4, noting that $\lambda > 4$ in (2.12), it holds

$$\widehat{\Omega} \stackrel{a.s.}{=} \frac{1}{T} \sum_{t=1}^T \xi_t \xi_t' + \mathcal{o}(T^{-1/2}).$$

The Central Limit Theorem for martingale differences by Brown and Eagleson (1971) is applicable when (2.12), (2.13) are satisfied, and implies that $\sqrt{T}(\widehat{\Omega} - \Omega)$ is asymptotically normal. Using the functional δ -method it is seen that this property is shared by $\widehat{\sigma}$ and $\widehat{\rho}$. ■

Theorem 4.2 *Suppose model (2.6) and Assumption 2.2, 2.3, 2.4, 2.5 are satisfied. Then, letting \odot denote the element wise (Hadamard) product,*

$$\left\{ \begin{array}{c} \widehat{\mathbb{U}}(u) \\ \widehat{\mathbb{V}}_1(u) \\ \widehat{\mathbb{V}}_2(u) \end{array} \right\} = \left[\begin{array}{c} 1 \\ \mathbf{f}\{\mathbb{F}^{-1}(u)\} \\ \frac{1}{2}\mathbb{F}^{-1}(u) \mathbf{f}\{\mathbb{F}^{-1}(u)\} \end{array} \right] \odot \int_0^1 \left[\begin{array}{c} 1_{(s \leq u)} \\ \mathbb{F}^{-1}(s) \\ \{\mathbb{F}^{-1}(s)\}^2 \end{array} \right] d\widehat{\mathbb{U}}(s) + \left\{ \begin{array}{c} 0 \\ 0 \\ \mathcal{O}_{\mathbb{P}}(1) \end{array} \right\},$$

which converges in distribution to $\mathbb{X}_{\mathbb{F}}(u)$ on $\mathbb{D}[0, 1]$.

Proof of Lemma 4.2. First, the stochastic integral representation in terms of $\widehat{\mathbb{U}}$ is established. For the first component this follows from the discussion in §2.2. For the other two terms, denote the empirical distribution of the innovations by $F_T(x) = T^{-1} \sum_{t=1}^T 1_{(\varepsilon_t \leq x)}$. As in §2.4 the sample and population moments of ε_t satisfy

$$\begin{aligned} \frac{1}{\sqrt{T}} \sum_{t=1}^T (\varepsilon_t^m - \mathbb{E}\varepsilon_t^m) &= \sqrt{T} \int_{\mathbf{R}} x^m d\{F_T(x) - F(x)\} \\ &= \int_{\mathbf{R}} x^m d\widehat{\mathbb{U}}\{F(x)\} = \int_0^1 \{F^{-1}(u)\}^m d\widehat{\mathbb{U}}(u), \quad (4.7) \end{aligned}$$

substituting $u = F(x)$. Thus the expression for $\widehat{\mathbb{V}}_1$, follows directly (4.4), which requires Assumption 2.3, whereas the expression for $\widehat{\mathbb{V}}_2$ follows from (4.5) and Lemma 4.1, requiring the Assumptions 2.2, 2.4.

Billingsley (1968, §13) shows that a continuous $\widehat{\mathbb{U}}$ converges to \mathbb{U} on $\mathbb{C}[0, 1]$. This requires Assumption 2.5. Since the limit is continuous, $\widehat{\mathbb{U}}$ converges to \mathbb{U} on $\mathbb{D}[0, 1]$. Further, due to the existence of fourth moments, see Assumption 2.4, the integrand in (4.7) is square integrable. Following Shorack and Wellner (1986, p.94) the integrator $\widehat{\mathbb{U}}$ can be replaced with the Brownian bridge \mathbb{U} in the limit. This gives the process \mathbb{X}_F in (2.14). ■

It is now argued that the component $\widehat{\mathbb{W}}_1$ vanishes. This result follows from the work of Koul (2002).

Theorem 4.3 *Suppose model (2.6) and Assumptions 2.2, 2.4, 2.7 are satisfied. Then*

$$\sup_{0 \leq u \leq 1} \left| \widehat{\mathbb{W}}_1(u) \right| = o_{\mathbb{P}}(1).$$

Proof. From Lemma 4.1 it follows that \widehat{a} is $O_{\mathbb{P}}(T^{-1/2})$. As pointed out by Rao and Sethuraman (1975) and Loynes (1980) a bound $b > 0$ can be found so $|\widehat{a}| < b$ with probability close to one. Thus it suffices to show

$$\sup_{\substack{x \in \mathbf{R} \\ |s| < b}} \frac{1}{\sqrt{T}} \sum_{t=1}^T \left[1_{\{\varepsilon_t \leq x(1+T^{-1/2}s)\}} - 1_{(\varepsilon_t \leq x)} - f(x) T^{-1/2} s \right].$$

This result follows from Corollary 2.3.2 of Koul (2002, p.59). A set of assumptions have to be checked. First, note that, in the notation of Koul (2002), $n = T$, $d_{ni} = T^{-1/2}$, $\mathbf{c}_{ni} = 0$, $X_{ni} = \varepsilon_t$, $H(x) = F_{ni} = F(x)$, $f_{ni}(x) = f(x)$. Since d_{ni} is uniform in t , the conditions to d_{ni} in **N1** and **N2** of Koul (2002, p.16) are trivially satisfied. Since $\mathbf{c}_{ni} = 0$ then the conditions to \mathbf{c}_{ni} in (2.3.6) and (2.3.7) of Koul (2002, p.52) are trivially satisfied. The conditions **F1**, **F2**, **F3** to f of Koul (2002, p.59) are satisfied by Assumption 2.7 as follows: **F1** requires uniform continuity of f , which is satisfied since f is differentiable and $\sup_{x \in \mathbf{R}} |f'(x)| < \infty$; **F2** requires f to be positive; **F3** requires $\sup_{x \in \mathbf{R}} |xf(x)| < \infty$. ■

Before looking at the components $\widehat{\mathbb{W}}_2$ and $\widehat{\mathbb{W}}_3$ it is useful to discuss the properties of the process $\widehat{z}_t = \widehat{a}x + \widehat{b}_t$ defined in (4.2).

Lemma 4.4 Suppose model (2.6) and Assumptions 2.2, 2.4 are satisfied. Then, if $g(T) \rightarrow \infty$ as $T \rightarrow \infty$,

(i) $\sigma \widehat{b}_t = (0, \widehat{\rho} - \rho) \xi_t + (1, -\widehat{\rho}) (\widehat{\theta} - \theta) S_{t-1}$,

(ii) $T^{-1/2} \sum_{t=1}^T \widehat{z}_t^2 = (1 + x^2) o_{\mathbb{P}}(1)$,

(iii) \widehat{b}_t can be written as $\widehat{b}_t = -\alpha_T \mathbf{z}_{Tt}$ where

(a) \mathbf{z}_{Tt} is \mathcal{G}_{t-1} -measurable as defined in Assumption 2.6,

(b) $\alpha_T = O_{\mathbb{P}}(1)$,

(c) $\sum_{t=1}^{T-g(T)} \mathbf{z}_{Tt} \mathbf{z}'_{Tt} = O_{\mathbb{P}}(1)$ for any function g so $g(T)/\log T \rightarrow \infty$.

Proof of Lemma 4.4. (i): rewrite \widehat{b}_t in terms of the parameters of the vector autoregression (2.6). By definition $\sigma \widehat{b}_t = \sigma \varepsilon_t - \widehat{\sigma} \widehat{\varepsilon}_t$, where $\sigma \varepsilon_t = (1, -\rho) \xi_t$ and $\widehat{\sigma} \widehat{\varepsilon}_t = (1, -\widehat{\rho}) \widehat{\xi}_t$, and ρ is either 0 or $\Omega_{yz} \Omega_{zz}^{-1}$. Adding and subtracting $(1, -\widehat{\rho}) \xi_t$ gives

$$\sigma b_t = \sigma \varepsilon_t - \widehat{\sigma} \widehat{\varepsilon}_t = (1, -\rho) \xi_t - (1, -\widehat{\rho}) \widehat{\xi}_t = (0, \widehat{\rho} - \rho) \xi_t - (1, -\widehat{\rho}) (\widehat{\xi}_t - \xi_t).$$

Writing the vector autoregression (2.6) in companion form

$$X_t = \theta S_{t-1} + \xi_t \quad \text{where} \quad \theta = (A_1, \dots, A_k, \mu), \quad S_{t-1} = (X'_{t-1}, \dots, X'_{t-k}, D'_{t-1})',$$

and in particular $\widehat{\xi}_t - \xi_t = -(\widehat{\theta} - \theta) S_{t-1}$ leads to the desired expression.

(ii): The least squares estimator for θ satisfies $\widehat{\theta} - \theta = M_{\xi S} M_{SS}^{-1}$ where

$$M_{\xi \xi} = \sum_{t=1}^T \xi_t \xi_t', \quad M_{\xi S} = \sum_{t=1}^T \xi_t S'_{t-1}, \quad M_{SS} = \sum_{t=1}^T S_{t-1} S'_{t-1}.$$

Using the inequality $(x + y)^2 \leq 2(x^2 + y^2)$ both to \widehat{z}_t^2 and to $(\sigma \widehat{b}_t)^2$ shows

$$\begin{aligned} \frac{1}{4} \sum_{t=1}^T (\sigma \widehat{z}_t)^2 &\leq \frac{T}{2} (\sigma \widehat{a} x)^2 + \frac{1}{2} \sum_{t=1}^T (\sigma \widehat{b}_t)^2 \\ &\leq T (\sigma \widehat{a} x)^2 + (0, \widehat{\rho} - \rho) M_{\xi \xi} (0, \widehat{\rho} - \rho)' + (1, -\widehat{\rho}) M_{\xi S} M_{SS}^{-1} M_{S \xi} (1, -\widehat{\rho})'. \end{aligned}$$

From Lemma 4.1, it follows that \widehat{a} and $\widehat{\rho} - \rho$ are both $O_{\mathbb{P}}(T^{-1/2})$. Further, Nielsen (2005, Theorem 2.4, 6.1) shows that $M_{\xi S} M_{SS}^{-1} M_{S \xi} = o_{\mathbb{P}}(T^{1/2})$ and $M_{\xi \xi} = O_{\mathbb{P}}(1)$. Both arguments use Assumptions 2.2, 2.4. Normalising by $T^{-1/2}$ gives the desired result.

(iii): From (i) it holds that $\sigma \widehat{b}_t = \sigma \alpha_T \mathbf{z}_{Tt}$, since, for instance in the presence of ρ ,

$$\alpha_T = - \left\{ (\widehat{\rho} - \rho), (1, -\widehat{\rho}) (\widehat{\theta} - \theta) \right\} N_T, \quad \mathbf{z}_{Tt} = N_T^{-1} \left\{ \begin{array}{c} (0, I_{p-1}) \xi_t \\ S_{t-1} \end{array} \right\},$$

for some normalisation matrix N_T . By construction \mathbf{z}_{Tt} is \mathcal{G}_{t-1} -measurable, showing (a). Note, that (ii) implies that $\sum_{t=1}^T \alpha_T \mathbf{z}_{Tt} \mathbf{z}'_{Tt} \alpha_T' = O_{\mathbb{P}}(T^{1/2})$, which is not quite sufficient for (b), (c). Thus, use that a matrix M exists so

$$M S_t = \begin{pmatrix} U_t \\ Q_t \\ W_t \end{pmatrix} = \begin{pmatrix} \mathbf{U} & 0 & 0 \\ 0 & \mathbf{Q} & 0 \\ 0 & 0 & \mathbf{W} \end{pmatrix} \begin{pmatrix} U_{t-1} \\ Q_{t-1} \\ W_{t-1} \end{pmatrix} = \begin{pmatrix} e_{U,t} \\ e_{Q,t} \\ e_{W,t} \end{pmatrix},$$

where the absolute values of the eigenvalues of \mathbf{U} , \mathbf{Q} , \mathbf{W} are less than one, equal to one, and greater than one, respectively, see Nielsen (2005, §3). The deterministic components are therefore included in the Q_t process. Using Theorems 2.4, 6.2 and Table 1 of Nielsen (2005) it follows that

$$\begin{pmatrix} \mathbf{M}_{\xi\xi} & \mathbf{M}_{\xi S} \\ \mathbf{M}_{S\xi} & \mathbf{M}_{SS} \end{pmatrix} = \text{diag}(\mathbf{M}_{\xi\xi}, \mathbf{M}_{UU}, \mathbf{M}_{QQ}, \mathbf{M}_{WW}) \{1 + o_{\mathbb{P}}(1)\}. \quad (4.8)$$

Therefore, write

$$\sigma b_t = -\sigma(\boldsymbol{\alpha}_{\xi,T} \mathbf{z}_{\xi,Tt}, \boldsymbol{\alpha}_{U,T} \mathbf{z}_{U,Tt}, \boldsymbol{\alpha}_{Q,T} \mathbf{z}_{Q,Tt}, \boldsymbol{\alpha}_{W,T} \mathbf{z}_{W,Tt}) \{1 + o_{\mathbb{P}}(1)\},$$

where, for $t \leq T - g(T)$ and some normalisation matrix $N_{Q,T}$,

$$\begin{aligned} \sigma \boldsymbol{\alpha}_{\xi,t} &= -(\widehat{\rho} - \rho) T^{1/2}, & \mathbf{z}_{\xi,Tt} &= T^{-1/2} (0, I_{p-1}) \xi_t, \\ \sigma \boldsymbol{\alpha}_{U,t} &= -(1, -\widehat{\rho}) \mathbf{M}_{\xi U} \mathbf{M}_{UU}^{-1} T^{1/2}, & \mathbf{z}_{U,Tt} &= T^{-1/2} U_{t-1}, \\ \sigma \boldsymbol{\alpha}_{Q,t} &= -(1, -\widehat{\rho}) \mathbf{M}_{\xi Q} \mathbf{M}_{QQ}^{-1} N_{Q,T}, & \mathbf{z}_{Q,Tt} &= N_{Q,T}^{-1} Q_{t-1}, \\ \sigma \boldsymbol{\alpha}_{W,t} &= -(1, -\widehat{\rho}) \mathbf{M}_{\xi W} \mathbf{M}_{WW}^{-1} \mathbf{W}^{T-g(T)}, & \mathbf{z}_{W,Tt} &= \mathbf{W}^{g(T)-T} W_{t-1}. \end{aligned}$$

For (b) it needs to be argued that each $\boldsymbol{\alpha}$ -term is $O_{\mathbb{P}}(1)$. For (c) it suffices to show that each \mathbf{z} -term has sums of squares that are $O_{\mathbb{P}}(1)$, and that the cross-products are $o_{\mathbb{P}}(1)$. For each of the following arguments it suffices that Assumptions 2.2, 2.4 are satisfied.

The ξ terms. Lemma 4.1 shows that $\boldsymbol{\alpha}_{\xi,T} = O_{\mathbb{P}}(1)$. Nielsen (2005, Theorem 6.2) shows that $\sum_{t=1}^T \mathbf{z}_{\xi,Tt} \mathbf{z}'_{\xi,Tt}$ is $O_{\mathbb{P}}(1)$.

The U term. The Central Limit Theorem for martingale differences by Brown and Eagleson (1971) implies that $\mathbf{M}_{\xi U} \mathbf{M}_{UU}^{-1/2} = O_{\mathbb{P}}(1)$. Nielsen (2005, Theorem 6.2) shows that $T^{-1} \mathbf{M}_{UU}$ has a positive definite limit, *a.s.* Thus $\boldsymbol{\alpha}_{U,T}$, $\sum_{t=1}^T \mathbf{z}_{U,Tt} \mathbf{z}'_{U,Tt}$ are both $O_{\mathbb{P}}(1)$.

The Q term. Using standard unit root weak convergence arguments as in Chan and Wei (1988) and Chan (1989) it can be shown that a normalisation matrix $N_{Q,T}$ exists so $\mathbf{M}_{\xi Q} N_{Q,T}$ and $N_{Q,T}^{-1} \sum_{t=1}^T Q_{t-1} Q'_{t-1} (N_{Q,T}^{-1})'$ are $O_{\mathbb{P}}(1)$, and the latter is positive definite *a.s.* Thus $\boldsymbol{\alpha}_{Q,T}$, $\sum_{t=1}^T \mathbf{z}_{Q,Tt} \mathbf{z}'_{Q,Tt}$ have the desired properties. Indeed Chan and Wei (1988), consider the univariate case, where $p = 1$ and $X_t = Y_t$, without deterministic terms, whereas Chan (1989) include deterministic terms for that case. The idea of Chan and Wei (1988) is first to show in their Theorem 2.2 that $T^{-1/2} \sum_{t=1}^{\text{int}(Tu)} \boldsymbol{\Lambda}_t \varepsilon_t$ converges to vector of independent standard Brownian motions, when defining

$$\boldsymbol{\Lambda}_t = \{1, (-1)^t, \sqrt{2} \sin(t\theta_1), \sqrt{2} \cos(t\theta_1), \dots, \sqrt{2} \sin(t\theta_l), \sqrt{2} \cos(t\theta_l)\},$$

for $\theta_k \in (0, \pi)$ so $\theta_k \neq \theta_j$ if $k \neq j$. This results is easily generalised to a multivariate result using the Cramér-Wold device, see Billingsley (1968). Next, existence of a normalisation matrix $N_{Q,T}$ and convergence of $T^{1/2} N_{Q,T}^{-1} Q_t$ and $N_{Q,T}^{-1} \sum_{t=1}^T Q_{t-1} Q'_{t-1} (N_{Q,T}^{-1})'$ then follows from the Continuous Mapping Theorem (see Chan and Wei, 1988, §3), which is therefore easily generalisable. For $\mathbf{M}_{\xi Q} N_{Q,T}$, convergence of $\mathbf{M}_{\xi Q} N_{Q,T}$ jointly with that of $T^{1/2} N_{Q,T}^{-1} Q_t$ is shown by a Skorokhod

embedding result when \mathbf{Q} is univariate (Chan and Wei, 1988, Theorem 2.4(ii)). This is easy to generalise when \mathbf{Q} is multivariate with simple roots (Chan and Wei, 1988, Remark to Theorem 2.4). For general \mathbf{Q} the Continuous Mapping Theorem is used (Chan and Wei, 1988, Theorem 2.4(i)), which again is easily generalisable. Finally, the positive definiteness of $N_{Q,T}^{-1} \sum_{t=1}^T Q_{t-1} Q'_{t-1} (N_{Q,T}^{-1})'$ is shown (Chan and Wei, 1988, Lemma 3.1.1), which can be generalised using a Cramér-Wold-type argument.

The W term. Nielsen (2005, Theorem 2.4, Corollary 7.2) shows $M_{\xi W} M_{WW}^{-1/2} = o(T^{1/4})$, *a.s.* and that $\mathbf{W}^{-T} M_{WW} (\mathbf{W}^{-T})'$ is convergent with positive definite limiting points. In particular $\sum_{t=1}^T \mathbf{z}_{W,Tt} \mathbf{z}'_{W,Tt} = \mathbf{W}^{g(T)-T} \sum_{t=1}^{T-g(T)} W_{t-1} W'_{t-1} (\mathbf{W}^{g(T)-T})'$ is convergent with positive definite limiting points, while $-\sigma \boldsymbol{\alpha}_{W,T} = o(T^{1/4} \mathbf{W}^{-g(T)})$ *a.s.*, which again is $o(1)$ for any $g(T)$ so $g(T)/\log T \rightarrow \infty$.

Cross terms. These all vanish. This follows directly from (4.8). ■

Theorem 4.5 *Suppose model (2.6) and Assumptions 2.2, 2.4, 2.7 are satisfied. Then*

$$\sup_{0 \leq u \leq 1} \left| \widehat{\mathbb{W}}_2(u) \right| = o_{\mathbf{P}}(1).$$

Proof of Theorem 4.5. At first $\widehat{\mathbb{W}}_2(u)$ is written as an integral

$$\widehat{\mathbb{W}}_2(u) = \frac{1}{\sqrt{T}} \sum_{t=1}^T \{F(x + \widehat{z}_t) - F(x) - f(x) \widehat{z}_t\} = \frac{1}{\sqrt{T}} \sum_{t=1}^T \int_x^{x+\widehat{z}_t} \{f(y) - f(x)\} dy.$$

By the triangle inequality

$$\left| \widehat{\mathbb{W}}_2(u) \right| \leq \frac{1}{\sqrt{T}} \sum_{t=1}^T \int_x^{x+\widehat{z}_t} |f(y) - f(x)| dy.$$

The integrand can be bounded by its maximum, so

$$\left| \widehat{\mathbb{W}}_2(u) \right| \leq \frac{1}{\sqrt{T}} \sum_{t=1}^T |\widehat{z}_t| \max_{|h| \leq |\widehat{z}_t|} |f(x+h) - f(x)|.$$

The mean value theorem then implies a further bound

$$\left| \widehat{\mathbb{W}}_2(u) \right| \leq \frac{2}{\sqrt{T}} \sum_{t=1}^T |\widehat{z}_t|^2 \max_{|h| \leq |\widehat{z}_t|} |f'(x+h)|.$$

Taking the maximum over the entire real axis, and using Lemma 4.4(ii), which requires Assumptions 2.2, 2.4, gives

$$\left| \widehat{\mathbb{W}}_2(u) \right| \leq o_{\mathbf{P}}(1) \sup_{x \in \mathbf{R}} |(1+x^2) f'(x)|,$$

which follows from Assumption 2.7. ■

The asymptotic uniform linearity property, that $\widehat{\mathbb{W}}_3$ vanishes, is now proved. Two ideas of Lee and Wei (1999) are used. First, to deal with the issue that for explosive component $W_T W_T'$ and $\sum_{t=1}^T W_t W_t'$, the largest and the smallest components are treated separately. Lee and Wei do actually not consider explosive and non-explosive components jointly, but the joint evaluation turns out to not to pose any problems. Secondly, Theorem 2.2 of Lee and Wei gives an asymptotic uniform linearity property for triangular arrays, which can be used here.

Theorem 4.6 *Suppose model (2.6) and Assumptions 2.2, 2.4, 2.5, 2.6, 2.7 are satisfied. Then*

$$\sup_{0 \leq u \leq 1} \left| \widehat{\mathbb{W}}_3(u) \right| = o_P(1).$$

Proof of Theorem 4.6. Some notation is needed. Let $u = F(x)$ and, recalling $\widehat{a}_t, \widehat{b}_t$ defined in (4.2), let $\widehat{x} = x(1 + \widehat{a}_t)$. Define

$$\mathbf{w}(t, \widehat{x}) = 1_{(\varepsilon_t \leq \widehat{x} + \widehat{b}_t)} - 1_{(\varepsilon_t \leq \widehat{x})} - F(\widehat{x} + \widehat{b}_t) + F(\widehat{x}).$$

Decompose $\widehat{\mathbb{W}}_3 = \widehat{\mathbb{W}}_{3,1} + \widehat{\mathbb{W}}_{3,2}$ where

$$\widehat{\mathbb{W}}_{3,1}(u) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T-g(T)} \mathbf{w}(t, \widehat{x}), \quad \widehat{\mathbb{W}}_{3,2}(\widehat{x}) = \frac{1}{\sqrt{T}} \sum_{t=T-g(T)+1}^T \mathbf{w}(t, \widehat{x}),$$

for some function $g(T)$ chosen so $g(T)/\sqrt{T} \rightarrow 0$ and $g(T)/\log T \rightarrow \infty$.

Analysis of $\widehat{\mathbb{W}}_{3,2}$. It is immediately seen that $|\mathbf{w}(t, u)| \leq 2$, so

$$\sup_{0 \leq u \leq 1} \left| \widehat{\mathbb{W}}_{3,2}(u) \right| \leq \frac{1}{\sqrt{T}} \sum_{t=T-g(T)+1}^T 2 = \frac{2g(T)}{\sqrt{T}} \rightarrow 0.$$

Analysis of $\widehat{\mathbb{W}}_{3,1}$. First, note that taking supremum over $x \in \mathbf{R}$ and over $\widehat{x} \in \mathbf{R}$ gives the same supremum so

$$\sup_{0 \leq u \leq 1} \left| \widehat{\mathbb{W}}_{3,1}(u) \right| = \sup_{x \in \mathbf{R}} \left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T-g(T)} \mathbf{w}(t, \widehat{x}) \right| = \sup_{x \in \mathbf{R}} \left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T-g(T)} \mathbf{w}(t, x) \right|.$$

Secondly, this can be written as

$$\sup_{x \in \mathbf{R}} \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \left\{ 1_{(\varepsilon_t \leq x - \boldsymbol{\alpha}_T \mathbf{z}_{Tt})} - 1_{(\varepsilon_t \leq x)} - F(x - \boldsymbol{\alpha}_T \mathbf{z}_{Tt}) + F(x) \right\} \right|,$$

where $-\boldsymbol{\alpha}_T \mathbf{z}_{Tt}$ equals \widehat{b}_t for $t \leq T-g(T)$ and zero otherwise. According to Lemma 4.4(iii) the term \mathbf{z}_{Tt} can be chosen as a triangular array of variables measurable with respect to the filtration \mathcal{G}_{t-1} introduced in Assumption 2.6.

Thirdly, the desired result now follows if the conditions of Lee and Wei (1999, Corollary 2.1) can be established. First, ε_t are independent and identically distributed with distribution function F according to Assumption 2.5, and independent of \mathcal{G}_{t-1} by Assumption 2.6. The vectors α'_T and \mathbf{z}_{Tt} have a dimension not depending on T where \mathbf{z}_{Tt} is \mathcal{G}_{t-1} -measurable. Since the ε_t s have the same marginal distribution function F with uniformly bounded second derivatives by Assumption 2.7 then Lee and Wei's condition (2.11) is trivially satisfied. Finally α_T and $\sum_{t=1}^T \mathbf{z}_{Tt} \mathbf{z}'_{Tt}$ are both $O_P(1)$ according to Lemma 4.4(iii), requiring Assumption 2.2, 2.4 ■

4.3 Empirical moments

Proof of Theorem 2.9. The following steps are taken. First, it is argued that the sample moments converge in distribution. Secondly, the identities (2.17), (2.18) are proved. Thirdly, the asymptotic covariance matrix is derived.

First, by using Lemma 4.1 along with (4.7), (2.16) then

$$\begin{aligned}\sqrt{T} \left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1 \right) &= \int_0^1 \{F^{-1}(u)\}^2 d\widehat{\mathbb{U}}(u), \\ \sqrt{T} (\widehat{\mu}_m - \mu_m) &= \int_0^1 \{F^{-1}(u)\}^m d\widehat{\mathbb{F}}(u).\end{aligned}$$

Since $\widehat{\mathbb{U}}$ is one of the components of $\widehat{\mathbb{F}}$, see (4.3), the joint convergence follows from Theorem 2.8, see Shorack and Wellner (1986, p.92). The second set of expressions for the limiting distribution in (2.17), (2.18) then follows. The first set of expressions will be derived below. The asymptotic, mean-zero, normality follows directly from the remarks about stochastic integrals with respect to \mathbb{U} in §2.3.

Secondly, the identity (2.17) is a straight forward definition. For the identity (2.18) some more work is required. The process \mathbb{X}_F is expressed in terms of \mathbb{U} in (2.14) showing that

$$\int_0^1 \{F^{-1}(u)\}^m d\mathbb{X}_F(u) = J_m + I_1 J_1 + I_2 J_2 \quad (4.9)$$

where the uniform process \mathbb{U} enters the J_m terms:

$$J_m = \int_0^1 \{F^{-1}(u)\}^m d\mathbb{U}(u),$$

whereas the term I_1 and I_2 are standard integrals not involving \mathbb{U} , and which will be shown to equal:

$$\begin{aligned}I_1 &= \int_0^1 \{F^{-1}(u)\}^m d[\mathbf{f}\{F^{-1}(u)\}] = -m\mu_{m-1}, \\ I_2 &= \int_0^1 \{F^{-1}(u)\}^m d\left[\frac{1}{2}F^{-1}(u)\mathbf{f}\{F^{-1}(u)\}\right] = -\frac{m}{2}\mu_m.\end{aligned}$$

The integral I_1 : The integrator is rewritten using the chain rule and implicit differentiation as

$$d [\mathbf{f} \{ \mathbf{F}^{-1} (u) \}] = \frac{\mathbf{f}' \{ \mathbf{F}^{-1} (u) \}}{\mathbf{f} \{ \mathbf{F}^{-1} (u) \}} du.$$

Inserting this in I_1 gives

$$I_1 = \int_0^1 \{ \mathbf{F}^{-1} (u) \}^m \frac{\mathbf{f}' \{ \mathbf{F}^{-1} (u) \}}{\mathbf{f} \{ \mathbf{F}^{-1} (u) \}} du.$$

Substituting $u = \mathbf{F} (x)$ so $du = \mathbf{f} (x) dx$ gives

$$I_1 = \int_{\mathbf{R}} x^m \frac{\mathbf{f}' (x)}{\mathbf{f} (x)} \mathbf{f} (x) dx = \int_{\mathbf{R}} x^m \mathbf{f}' (x) dx.$$

Finally, by partial integration

$$I_1 = -m \int_{\mathbf{R}} x^{m-1} \mathbf{f} (x) dx = -m \mu_{m-1}.$$

The integral I_2 : In the same way

$$d \left[\frac{1}{2} \mathbf{F}^{-1} (u) \mathbf{f} \{ \mathbf{F}^{-1} (u) \} \right] = \frac{1}{2} \left[1 + \mathbf{F}^{-1} (u) \frac{\mathbf{f}' \{ \mathbf{F}^{-1} (u) \}}{\mathbf{f} \{ \mathbf{F}^{-1} (u) \}} \right] du.$$

Inserting this in I_2 , substituting $u = \mathbf{F} (x)$ and using partial integration

$$I_2 = \frac{1}{2} \int_{\mathbf{R}} x^m \left\{ 1 + x \frac{\mathbf{f}' (x)}{\mathbf{f} (x)} \right\} \mathbf{f} (x) dx = \frac{1}{2} \{ \mu_m - (m+1) \mu_m \} = -\frac{m}{2} \mu_m.$$

Finally the covariance matrix: The covariance follow from the formula (2.11). Thus, the expression for σ_{22} follows from

$$\int_0^1 h_2^2 (u) du = \int_0^1 \{ \mathbf{F}^{-1} (u) \}^4 du = \mu_4, \quad \int_0^1 h_2 (u) du = \mu_2 = 1.$$

The other expressions follow in a similar way. First,

$$\int_0^1 h_m (u) du = \mu_m - m \mu_{m-1} \mu_1 - \frac{m}{2} \mu_m \mu_2 = \mu_m \left(1 - \frac{m}{2} \right),$$

using $\mu_1 = 0$ and $\mu_2 = 1$, while

$$\int_0^1 h_m (u) h_2 (u) du = \mu_{m+2} - m \mu_{m-1} \mu_3 - \frac{m}{2} \mu_m \mu_4$$

cannot be reduced further. Together with the above expressions this leads to σ_{m2} . Finally, σ_{mn} can be derived using

$$\begin{aligned} \int_0^1 h_m (u) h_n (u) du &= \mu_{m+n} - n \mu_{m+1} \mu_{n-1} - \frac{n}{2} \mu_{m+2} \mu_n \\ &\quad - m \mu_{m-1} \mu_{n+1} + mn \mu_{m-1} \mu_{n-1} \mu_2 + \frac{mn}{2} \mu_{m-1} \mu_n \mu_3 \\ &\quad - \frac{m}{2} \mu_m \mu_{n+2} + \frac{mn}{2} \mu_m \mu_{n-1} \mu_3 + \frac{mn}{4} \mu_m \mu_n \mu_4. \end{aligned}$$

■

4.4 The Gaussian case

Proof of Theorem 2.9. It is first established, for $\Phi(x) = u$ that

$$\begin{aligned}\int_0^1 1_{(s \leq u)} \Phi^{-1}(s) ds &= -\varphi(x), \\ \int_0^1 1_{(s \leq u)} \{\Phi^{-1}(s)\}^2 ds &= -x\varphi(x) + u.\end{aligned}$$

To see this, note that $\varphi'(x) = -x\varphi(x)$ and $\varphi''(x) = (x^2 - 1)\varphi(x)$, and therefore by substitution $\Phi(t) = s$ it holds

$$\int_0^1 1_{(s \leq u)} \Phi^{-1}(s) ds = \int_{-\infty}^x t\varphi(t) dt = -\int_{-\infty}^x \varphi'(t) dt = -\varphi(x),$$

and likewise

$$\begin{aligned}\int_0^1 1_{(s \leq u)} \{\Phi^{-1}(s)\}^2 ds &= \int_{-\infty}^x \{(t^2 - 1) + 1\} \varphi(t) dt \\ &= \int_{-\infty}^x \{\varphi''(t) + \varphi(t)\} dt = \varphi'(x) + \Phi(x).\end{aligned}$$

The integrals involved in the expression (2.11) for the covariance of integrals with respect to the Brownian bridge can now be considered. Recalling the integral expressions for the Gaussian moments in (2.19) it holds

$$\varkappa_u = \int_0^1 \begin{pmatrix} 1_{(s \leq u)} \\ \Phi^{-1}(s) \\ \{\Phi^{-1}(s)\}^2 \end{pmatrix} ds = \begin{pmatrix} u \\ 0 \\ 1 \end{pmatrix},$$

and for $u = \Phi(x)$ and $v = \Phi(y)$ so $u \leq v$

$$\begin{aligned}\Sigma_{u,v} &= \int_0^1 \begin{pmatrix} 1_{(s \leq u)} \\ \Phi^{-1}(s) \\ \{\Phi^{-1}(s)\}^2 \end{pmatrix} \begin{pmatrix} 1_{(s \leq v)} \\ \Phi^{-1}(s) \\ \{\Phi^{-1}(s)\}^2 \end{pmatrix} ds \\ &= \begin{pmatrix} u & -\varphi(x) & u - x\varphi(x) \\ -\varphi(y) & 1 & 0 \\ u - y\varphi(y) & 0 & 3 \end{pmatrix}.\end{aligned}$$

The desired expression is then computed as

$$\left\{ 1, \varphi(x), \frac{1}{2}x\varphi(x) \right\} (\Sigma_{u,v} - \varkappa_u \varkappa'_v) \left\{ 1, \varphi(y), \frac{1}{2}y\varphi(y) \right\}'.$$

■

5 Acknowledgements

The numerical results were generated using Ox (see Doornik, 1999). The second author received financial support from ESRC grant RES-000-27-0179.

6 References

- Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23, 193-212.
- Billingsley (1968). *Convergence of Probability Measures*. New York: Wiley.
- Boldin (1981). Estimation of the distribution of noise in an autoregressive scheme. *Theory of Probability and its Applications*, 27, 866-871.
- Brown, B.M. and Eagleson, G.K. (1971). Martingale convergence to infinitely divisible laws with finite variance. *Transactions of the American Mathematical Society* 162, 449-453.
- Chan, N.H. and Wei, C.Z. (1988). Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes. *Annals of Statistics* 16, 367-401.
- Chan, N.H. (1989). Asymptotic inference for unstable autoregressive time series with drifts. *Journal of Statistical Planning and Inference* 23, 301-312.
- Čibisov, D.M. (1966). Some theorems on the limiting behavior of the empirical distribution function. *Selected Translations in Mathematical Statistics and Probability*, 6, 147-156.
- Doornik, J.A. (1999) *Object-oriented matrix programming using Ox*, 3rd ed. London: Timberlake Consultants Press.
- Fisher, R.A. (1930). The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London*, A130, 16-28.
- Jarque, C.M. and Bera, A.K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55, 163-172.
- Johansen, S. (1995). *Likelihood-based Inference for Cointegration*. Oxford: Oxford University Press.
- Johansen, S. (2000). A Bartlett correction factor for tests on the cointegrating relations. *Econometric Theory* 16, 740-778.
- Johnson, N.L., Kotz, S., and Balakrishnan, N. (1994) *Continuous Univariate Distributions*. New York: Wiley.
- Kilian, L. and Demiroglu, U. (2000). Residual-based tests for normality in autoregressions: Asymptotic theory and simulation evidence. *Journal of Business and Economic Statistics*, 18, 40-50.
- Koul, H.L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models*, 2nd edition. New York: Springer.

- Koul, H.L. and Leventhal, S. (1989). Weak convergence of the residual empirical process in explosive autoregression. *Annals of Statistics*, 17, 1784-1794.
- Lai, T.L. and Wei, C.Z. (1985). Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. In P.R. Krishnaiah (ed.), *Multivariate Analysis VI*, pp. 375-393. Amsterdam: Elsevier Science Publishers.
- Lauritzen, S.L. (2002). *Thiele: Pioneer in Statistics*. Oxford: Oxford University Press.
- Lee, S. and Wei, C.Z. (1999). On residual empirical processes of stochastic regression models with applications to time series. *Annals of Statistics*, 27, 237-261.
- Loynes, R.M. (1980). The empirical distribution function of residuals from generalized regression. *Annals of Statistics*, 8, 285-298.
- Nielsen, B. (1997). Bartlett correction of the unit root test in autoregressive models. *Biometrika* 84, 500-504.
- Nielsen, B. (2001). The Asymptotic Distribution of Unit Root Tests of Unstable Autoregressive Processes. *Econometrica*, 69, 211-219
- Nielsen, B. (2005). Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms. *Econometric Theory* 21, 534-561.
- Nielsen, B. (2006). Correlograms for non-stationary autoregressions. *Journal of the Royal Statistical Society*, B68, 707-720.
- Nielsen, B. (2007). Order determination in general vector autoregressions. To appear in Ho, H.-C., Ing, C.-K., and Lai, T.L. *Time Series and Related Topics: In Memory of Ching-Zong Wei* IMS Lecture Notes and Monograph Series, volume 52.
- Pearson, K. (1902). On the Mathematical Theory of Errors of Judgment, with Special Reference to the Personal Equation. *Philosophical Transactions of the Royal Society of London*, A198, 235-299.
- Pierce, D.A. (1985). Testing normality in autoregressive models. *Biometrika*, 72, 293-297.
- Rao, J.S. and Sethuraman, J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. *Annals of Statistics*, 3, 299-313.
- Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons.
Journal of the American Statistical Association, 69, 730-737.

Thiele, T.N. (1889). *Almindelig Iagttagelseslære: Sandsynlighedsregning of mindste Kvadraters Methode*. Copenhagen: C.A. Reitzel.