Concepts and tools for nonlinear time series modelling

Amendola, Alessandra and Francq, Christian

Department of Economics and Statistics, University of Salerno, EQUIPPE-GREMARS, Université Lille III

2009

Concepts and tools for nonlinear time series modelling

Alessandra Amendola

Department of Economics and Statistics, University of Salerno, 84084 Fisciano (SA), Italy

Christian Francq

Université Lille III, EQUIPPE-GREMARS Universités de Lille, BP 60149, 59653 Villeneuve d'Ascq cedex, France

Abstract

Tools and approaches are provided for nonlinear time series modelling in econometrics. A wide range of topics is covered, including probabilistic properties, statistical inference and computational methods. The focus is on the applications but the ideas of the mathematical arguments are also provided. Techniques and concepts are illustrated by various examples, Monte Carlo experiments and a real application. [1]

1 Introduction

For the modelling and prediction of data collected sequentially in time, practitioners possess a well-established methodology based on linear time series models. This is the so-called Box-Jenkins methodology, which consists in fitting autoregressive moving-average (ARMA) models by model selection and estimation followed by model criticism through significance tests and diagnostic checks on the adequacy of the fitted model (see *e.g.* the comprehensive book [29]). A univariate time series (Y_t) satisfies an ARMA model when

$$Y_t - \sum_{i=1}^{p} a_i Y_{t-i} = \nu + \epsilon_t - \sum_{i=1}^{q} b_i \epsilon_{t-i},$$

[1] This document is an extended version of : A. Amendola, C. Francq, Concepts and tools for nonlinear time series modelling, Handbook of Computational Econometrics, Edts: D. Belsley and E. Kontoghiorghes, Wiley (2009).

where the ϵ_t are errors terms. The popularity of these models is certainly due to their relatively simple mathematical tractability and also to the existence of computer software incorporating the above-mentioned Box-Jenkins methodology. The ARMA models appear however insufficient because they are not able to take into account important features of many observed data, such as the conditional heteroscedasticity of the financial times series. The autoregressive conditional heteroscedastic (ARCH) models have been introduced by Engle [51] to allow stationarity (in particular a time invariant unconditional variance) with time-varying conditional variance through an equation of the form

$$\epsilon_t = \sigma_t \eta_t, \quad \sigma_t^2 = \omega + \sum_{i=1}^{Q} \alpha_i \epsilon_{t-i}^2. \tag{1}$$

where η_t is an independent and identically distributed (iid) sequence of random variables with mean 0 and variance 1. Numerous other parametric specifications of the conditional variance have been proposed in the literature. The most widely used ARCH-type models are today the GARCH models of [21]. In these models the term $\sum_{i=1}^{P} \beta_i \sigma_{t-i}^2$ is added to the right-hand side of (1), which allows for long run effects of the shocks. The EGARCH of [97], the GJR-GARCH of [64], the APARCH of [50], the QGARCH of [107] or the TARCH of [128] are also widely employed to take into account the asymmetric impacts of the shocks on the volatility. Figure 1 compares the news impact curve (*i.e.* the function $\epsilon_t \mapsto \sigma_{t+1}^2$) of 3 different ARCH-type models (see [54] for details on the concept of news impact curve). Note that the TARCH and QARCH models allow good news (*i.e.* positive returns $\epsilon_t > 0$) and bad news (*i.e.* negative returns $\epsilon_t < 0$) to have a different impact on the future volatility σ_{t+1}, and that for the QARCH model the volatility is not minimal at 0. The stochastic volatility models of [110] have been introduced to allow a more flexible form, specifying σ_t itself as a stochastic process which is not only driven by past observable variables.

It should be noted that the specification of $\sigma_t^2 := \mathrm{Var}(\epsilon_t \mid \epsilon_u, u < t)$ does not affect the conditional mean $E(Y_t \mid Y_u,\, u < t)$, which in the simple AR(p) case is the linear function $\nu + \sum_{i=1}^{p} a_i Y_{t-i}$. A natural extension of the linear AR model is the general non linear model

$$Y_t = F_\theta(X_t, Y_{t-1}, \ldots, Y_{t-k}) + \epsilon_t, \tag{2}$$

where F_θ is a function which may be nonlinear and X_t is a vector of exogenous variables. One of the most popular non-linear model belonging to the general specification (2) is the self-exciting threshold autoregressive (SETAR) model (see [121] and the monograph [120]). A two-regime SETAR model is defined by

$$Y_t = \left\{ \nu^{(1)} + \sum_{i=1}^{p} a_i^{(1)} Y_{t-i} + \epsilon_t^{(1)} \right\} + \left\{ \nu^{(2)} + \sum_{i=1}^{p} a_i^{(2)} Y_{t-i} + \epsilon_t^{(2)} \right\} \mathbf{1}_{\{Y_{t-d} > r\}}, \tag{3}$$

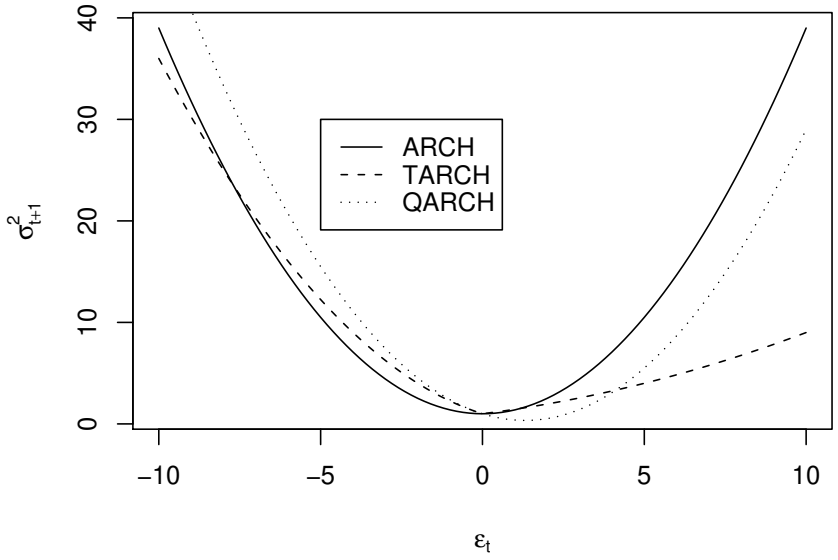


Figure 1. News impact curves of the ARCH(1) model $\epsilon_t = \sqrt{1 + 0.38\epsilon_{t-1}^2}\eta_t$, the TARCH(1) model $\epsilon_t = (1 - 0.5\epsilon_{t-1}^- + 0.2\epsilon_{t-1}^+)\eta_t$, and the QARCH(1) model $\epsilon_t = \sqrt{1 + 0.38\epsilon_{t-1}^2 - \epsilon_{t-1}}\eta_t$.

where $\mathbf{1}_A$ is the indicator function of the event A, r is the threshold parameter and d is the threshold delay. The explanation of the term "self-exciting" is that the dynamics switches from one to the other AR regime according to the past of (Y_t) itself. The formulation can be extended to the SETARMA model (see [119]), including a moving average component.

To model periodic phenomena it may be more natural to assume that the AR parameter changes deterministically over time. This leads to the so-called time-varying models (see [13], [18] and the references therein).

It has been suggested (see [111] and the references therein) to modify (3) in order to allow smooth transitions between the regimes. The transition is driven by a transition function $G(\gamma, c, s_t)$ in which (s_t) is a transition variable, which is not necessarily of the form $s_t = Y_{t-d}$, but can be a more general function of the variables $X_t, Y_{t-1}, \ldots, Y_{t-p}$. An example of transition function is the logistic function

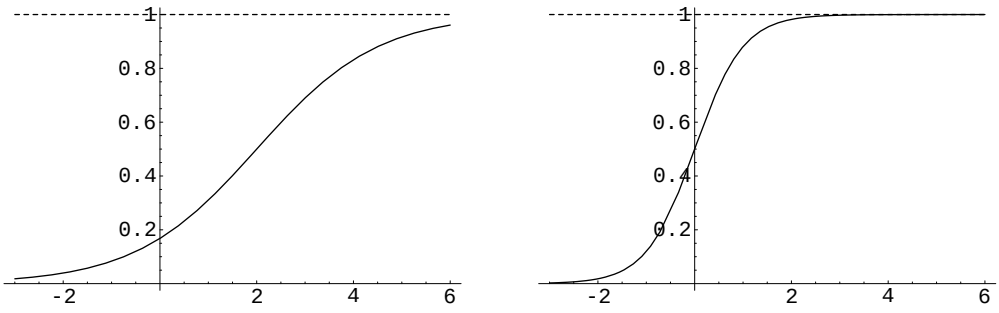
$$G(\gamma, c, s_t) = \frac{1}{1 + \exp\{-\gamma(s_t - c)\}},$$

where $\gamma > 0$ is a slope parameter and c is a location parameter (see Figure 2). In its simplest form the smooth transition regression (STR) can be written as

$$Y_t = \beta_1' W_t + \{\beta_2' W_t\} G(\gamma, c, s_t) + \epsilon_t$$

where $W_t = (1, Y_{t-1}, \ldots, Y_{t-p}, X_t')'$.

The exponential autoregressive (EXPAR) models introduced by [70] are de-



$x \to G(0.8, 2, x)$ $x \to G(2, 0, x)$

Figure 2. Logistic smooth transition function $x \to G(\gamma, c, x)$.

fined by

$$Y_t = \sum_{i=1}^{p} \left\{ a_i + b_i \exp(-\gamma Y_{t-1}^2) \right\} Y_{t-i} + \epsilon_t$$

with $\gamma > 0$. These models are able to account for limit cycles and can be considered as particular smooth transition autoregressive (STAR) models. Indeed, the dynamics is close to that of an AR(p) model with coefficients $(a_1, \ldots, a_p)$ when Y_{t-1}^2 is large, and with coefficients tending to $(a_1 + b_1, \ldots, a_p + b_p)$ when Y_{t-1}^2 decreases. The EXPAR models can also be viewed as particular cases of more general random coefficient autoregressive (RCA) models (see [98]) or functional autoregressive (FAR) models (see [33]).

Another class of models with regime changes is obtained when the transition variable is an unobserved random process. If one assumes the existence of several AR regimes, and that the dynamics switches from one regime to another regime according to a non-observed Markov chain, we obtain a Markov-switching model of the form

$$Y_t = c(\Delta_t) + \sum_{i=1}^{p} a_i(\Delta_t) Y_{t-i} + \sigma_t(\Delta_t)\epsilon_t, \tag{4}$$

where (Δ_t) is a Markov chain with finite state-space $\mathcal{E} = \{1, 2, ..., d\}$. An increasing interest for this class has been shown in the econometric literature since the seminal paper by Hamilton [72] who introduced a business cycle model of the form $Y_t = c(\Delta_t) + X_t$ where X_t is an AR(p) model. These models extend the class of hidden Markov models (HMM), in which the observations are assumed to be independent conditional on the hidden Markov chain Δ_t. The HMM have been introduced by [16] and have found numerous applications, for instance in speech recognition (see *e.g.* [103]).

The bilinear models (see [67] and [109]) constitute another very popular class

of nonlinear models, defined by

$$Y_t = \epsilon_t + \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=1}^{q} b_j \epsilon_{t-j} + \sum_{i=1}^{P}\sum_{j=1}^{Q} c_{ij} Y_{t-i}\epsilon_{t-j}. \tag{5}$$

When $P = Q = 0$, the product terms $Y_{t-i}\epsilon_{t-j}$ disappear and one obtains a linear ARMA(p, q) model.

Tong [120] first suggested to combine a nonlinear model for the conditional mean with a model for the conditional variance. For this purpose, models such as TAR-ARCH and BL-ARCH have been used in practical application. The empirical evidence on the presence of asymmetry, both in the level and in the conditional variance, leads to DTARCH (Double Threshold Autoregressive Conditionally Heteroscedastic) models, first proposed by Li and Li [85] and further investigated in [87]. A DTARCH model is given by

$$\begin{cases} Y_t &= \sum_{j=1}^{k} (\nu^{(j)} + \sum_{i=1}^{p_j} a_i^{(j)} Y_{t-i} + \epsilon_t) \mathbf{1}_{\{r_{j-1} < Y_{t-d} \leq r_j\}} \\ \epsilon_t &= \sigma_t \eta_t \\ \sigma_t^2 &= \sum_{j=1}^{k} (\omega^{(j)} + \sum_{i=1}^{Q_j} \alpha_i^{(j)} \epsilon_{t-i}^2) \mathbf{1}_{\{c_{j-1} < \epsilon_{t-s} \leq c_j\}} \end{cases}$$

where the threshold values, r_j and c_j, are such that $r_0 < r_1 < ... < r_k$, $r_0 = -\infty$, $r_k = +\infty$, and $c_0 < c_1 < ... < c_k$, $c_0 = -\infty$, $c_k = +\infty$; d and s are the delay parameters for the conditional mean and the conditional variance.

There are many other classes of parametric nonlinear models. Non parametric methods (see [75]) and semiparametric models (see [55]) are also very useful to forecast and analyze nonlinear time series.

The aim of the paper is to examine some of the key issues in nonlinear time series analysis, limiting ourselves to univariate and stationary series. A large number of papers and books are also devoted to this topic (see [39], [55], [62], [69], [102], [108], [112], [113], [120], [123] and the references therein). The approach chosen in the present paper is to concentrate on tools and methods, rather than on models themselves. This choice was done for two reasons. First the number of imaginable nonlinear models is virtually infinite, whereas fundamental concepts like stationarity and ergodicity are of interest for all these models. Second, comprehensive reviews of the nonlinear models used in financial and macroeconomic time series are already contained in the above mentioned works.

We tried to make the text as self-consistent as possible, and to give the main ideas of the mathematical arguments whenever that was not too difficult. The concepts are illustrated by original examples and Monte Carlo simulations. This paper is intended for a broad public of researchers in statistics or

econometrics, coming either from theoretical domains or from applied areas, but who are not familiar with all the aspects of nonlinear time series modelling. For example, a probabilist will learn nothing from the section devoted to the probabilistic tools, but could be interested by the importance of the geometric ergodicity for the Markov chains generated by the MCMC methods. Conversely, an applied analyst using MCMC algorithms could be glad to see that, at least for some simple models, it is possible to find mild conditions for geometric ergodicity. This is why we decided not to concentrate on only one aspect of non linear time series analysis, such as for instance the computational methods.

The rest of this chapter is organized as follows. In Section 2 we give the definitions of linear and nonlinear data generating processes, and we discuss the concept of model. We argue that linear and non-linear models are not incompatible and can often be complementary. Section 3 presents parametric and nonparametric tests statistics that can discriminate linear series from nonlinear ones. Section 4 provides tools for deriving the main probabilistic properties of nonlinear time series models. Section 5 is devoted to the statistical inference. Section 6 concerns forecasting issues. Section 7 is devoted to numerical and computational aspects. Section 8 concludes.

It is not necessary to read sequentially the different sections, according to the topic, a "nonlinear" reading is recommended.

2 Nonlinear data generating processes and linear models

In this section we make the distinction between a data generating process (DGP) and a model. A model is generally insufficient to determine the whole distribution of the data, but is simply an equation which can indeed be satisfied by processes of all types.

2.1 Linear and nonlinear processes

It is not so obvious to define the notion of linear and nonlinear process. These concepts are sometimes used in a different sense from the one adopted here. The process $(X_t)_{t\in\mathbb{Z}}$ is said to be a linear process with mean 0 if for all $t \in \mathbb{Z}$

$$X_t = \eta_t + \sum_{i=1}^{\infty} \psi_i \eta_{t-i}, \quad \sum_{i=1}^{\infty} \psi_i^2 < \infty, \qquad (\eta_t) \text{ is IID}(0, \sigma^2), \tag{6}$$

and

$$X_t = \eta_t + \sum_{i=1}^{\infty} \pi_i X_{t-i}, \quad \sum_{i=1}^{\infty} \pi_i^2 < \infty, \tag{7}$$

where $\mathrm{IID}(0, \sigma^2)$ denotes an iid sequence of random variables with mean 0 and common variance $\sigma^2 > 0$. Such a sequence is sometimes called a *strong white noise*. A *weak white noise* is a stationary sequence of centered and uncorrelated random variables with common variance $\sigma^2 > 0$, and will be denoted by $\mathrm{WN}(0, \sigma^2)$. Obviously a strong white noise is also a weak white noise, because independence entails uncorrelatedness, but the reverse is not true. We will see that the distinction between strong and weak white noises is fundamental in nonlinear time series analysis.

Under standard assumptions (see Examples 1 and 2 below), the ARMA processes, and also processes with long memory, satisfy (6)-(7). It is shown in [27] and [34] that the "two-sided" linear representation

$$X_t = \eta_t + \sum_{i=-\infty}^{\infty} \psi_i \eta_{t-i}, \quad \sum_{i=-\infty}^{\infty} \psi_i^2 < \infty, \qquad (\eta_t) \sim \mathrm{IID}(0, \sigma^2), \tag{8}$$

is essentially unique when (X_t) is not gaussian and when the spectral density of (X_t) is positive almost everywhere (more precisely if $X_t = \eta_t^* + \sum_{i=-\infty}^{\infty} \psi_i^* \eta_{t-i}^*$, with $\sum_{i==-\infty}^{\infty} \psi_i^{*2} < \infty$, and $(\eta_t^*) \sim \mathrm{IID}(0, \sigma^{*2})$, then $\eta_t^* = c\eta_{t-s}$ and $\psi_i^* = \psi_{i+s}/c$ for some $s \in \mathbb{Z}$ and some $c > 0$). Thus, when the spectral density of (X_t) is positive almost everywhere, *the linear representation (6) is unique, except in the gaussian case.*

Example 1 *Consider a MA(1) process of the form $X_t = \eta_t - b\eta_{t-1}$ with $|b| > 1$ and η_t $\mathrm{IID}(0, \sigma^2)$. This representation is said to be noninvertible because η_t can not be expressed as a function of $\{X_u,\ u \le t\}$. Instead one has the "anticipative" representation $\eta_t = -\sum_{i\ge 1} b^{-i} X_{t+i}$. It is easy to see that $\epsilon_t := \sum_{i\ge 0} b^{-i} X_{t-i} = \eta_t + \sum_{i\ge 1} b^{-i}(1-b^2)\eta_{t-i}$ is a $\mathrm{WN}(0, b^2\sigma^2)$. Thus (X_t) also satisfies the "invertible" MA(1) representation $X_t = \epsilon_t - b^{-1}\epsilon_{t-1}$. When (X_t) is Gaussian, the processes (η_t) and (ϵ_t) are also Gaussian, and ϵ_t is $\mathrm{IID}(0, b^2\sigma^2)$ because for gaussian processes the two concepts of white noise coincide. A nontrivial consequence of the uniqueness of (6) in the non-Gaussian case is that the ϵ_t's are not independent when (X_t) is not Gaussian.*

Example 2 *As in the previous example, it can be shown that the non causal AR(1) equation*

$$X_t - aX_{t-1} = \eta_t, \qquad |a| > 1, \qquad (\eta_t) \sim \mathrm{IID}(0, \sigma^2)$$

admits an anticipative stationary solution. This solution also satisfies the causal AR(1) equation $X_t - a^{-1}X_{t-1} = \epsilon_t$, where ϵ_t is a weak white noise. Except in the Gaussian case, this noise is not strong.

We will say that a process (X_t) is nonlinear if (6)-(7) does not hold. With this definition, some noncausal ARMA processes (as in Example 2) are considered

as nonlinear. This is also the case for the so-called all-pass models, which are causal ARMA models in which the roots of the AR polynomial are the reciprocals of the roots of the MA polynomial (see [28] and the references therein for details about all-pass models). The following example corresponds to the simplest all-pass model.

Example 3 *Consider the process defined by*

$$X_t - aX_{t-1} = \eta_t - \frac{1}{a}\eta_{t-1}, \quad |a| < 1 \quad (\eta_t) \sim \mathrm{IID}(0, \sigma^2). \tag{9}$$

It is easy to see that the spectral density of X_t is constant, equal to $\sigma^2/(a^2 2\pi)$. Thus X_t is $\mathrm{WN}(0, \sigma^2/a^2)$ and, in view of the previous arguments, X_t is $\mathrm{IID}(0, \sigma^2/a^2)$ if and only if η_t is Gaussian. Figure 3 shows that, when η_t is not gaussian, the simulated trajectories of (X_t) may share common features with financial returns : uncorrelation of the observed values, but strong correlations of the squares or of the absolute values, and volatility clustering. Thus the non-Gaussian all-pass models are not strong white noises, though they are weak white noises.

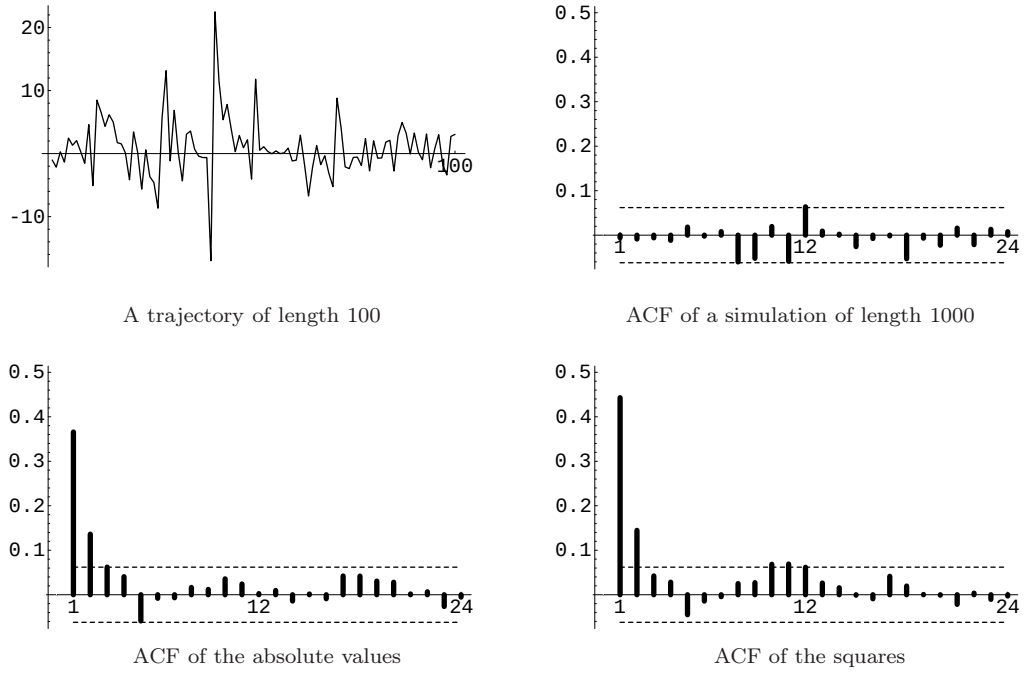


Figure 3. Simulation of the all-pass model (9) where $a = 0.5$ and η_t follows a Student distribution with 3 degrees of freedom, sample autocorrelations of X_t, of $|X_t|$, and of X_t^2 for $t = 1, \ldots, n = 1000$.

Condition (8) is sometimes taken as definition of a linear process. In this case, the all-pass models are considered as linear. This does not seem desirable because, as we have seen, the behavior of the all-pass models (uncorrelatedness of the observations and correlation of their squares) is often considered as a typical nonlinearity feature.

Wold [126] has shown that any purely non deterministic, second-order stationary process admits an infinite MA representation of the form

$$X_t = \epsilon_t + \sum_{i=1}^{\infty} \psi_i \epsilon_{t-i}, \quad \sum_{i=1}^{\infty} \psi_i^2 < \infty, \qquad (\epsilon_t) \sim \mathrm{WN}(0, \sigma^2), \tag{10}$$

where (ϵ_t) is the linear innovation process of (X_t). Comparing (6) and (10), one can see that

(X_t) is linear when its innovation process (ϵ_t) is a strong white noise. (11)

Note that (10) is just a representation, a model, but is not sufficient to entirely define the process (X_t). Several linear or nonlinear DGP may admit the same linear model of the form (10).

Example 4 *Consider the simple bilinear process $X_t = \eta_t + bX_{t-1}\eta_{t-2}$, where $b^2 < 1$ and (η_t) is* $\mathrm{IID}(0,1)$. *The stationary solution writes*

$$X_t = \eta_t + b\eta_{t-1}\eta_{t-2} + \sum_{k=2}^{\infty} b^k \eta_{t-2}\eta_{t-3}\cdots\eta_{t-k+1}\eta_{t-k}^2\eta_{t-k-1}.$$

One can easily check that $E(X_t) = 0$, $\gamma(0) = 1 + b^2 + \mu_4 b^4/(1-b^2)$, $\gamma(1) = b^3\mu_3$, $\gamma(2) = 0$, $\gamma(3) = b^2$, and $\gamma(h) = 0$ for all $h > 3$, where $\mu_3 = E\eta_t^3$, $\mu_4 = E\eta_t^4$ and $\gamma(h) = Cov(X_t, X_{t-h})$. When $\mu_3 = 0$, X_t satisfies the MA(3) representation $X_t = \epsilon_t + \alpha_3\epsilon_{t-3}$, where α_3 is obtained by solving $\alpha_3/(1+\alpha_3^2) = b^2(1-b^2)/(1-b^4+b^4\mu_4)$, $|\alpha_3| < 1$. Note that from this weak MA(3) representation the optimal linear prediction of X_t given its past takes the form $\alpha_3 X_{t-3} - \alpha_3^2 X_{t-6} + \alpha_3^3 X_{t-9} + \cdots$ whereas the optimal prediction is $bX_{t-1}X_{t-2} - b^2 X_{t-1}X_{t-3}X_{t-4} + b^3 X_{t-1}X_{t-3}X_{t-5}X_{t-6} + \cdots$, provided this expansion exists.

Numerous other examples of nonlinear processes admitting weak ARMA representations can be found in [61] and the references therein.

3 Testing linearity

Most of the classes of nonlinear models encompass linear ones. This is clearly the case for the bilinear models (5). Thus, it may be argued that a bilinear model can always provide forecasts at least as good as those of an ARMA model. The problem is obviously that the parameters have to be estimated. This is why a simple model, though often incomplete, may outperform a more

complicated model. This leads the practitioner to adopt the principle of parsimony, by preferring the simplest models and rejecting the unnecessarily complicated specifications. Many nonlinear models are not identified when the DGP is linear. This is the case, for instance, with the SETAR model (3). When there is only one regime, one can take the threshold $r = +\infty$ and arbitrary values for the parameters $\nu^{(2)}$, $\sigma^{(2)}$ and $a_i^{(2)}$. It will be seen in Section 3.3 that such identifiability problems entail difficulties to determine the behavior of estimators of nonlinear models when the DGP is linear. As a consequence, a building-model strategy "from general to specific" (beginning with the estimation of a very large nonlinear model, followed by successive cancelation of non significant coefficients) is not always a good idea.

For these reasons, one can recommend to begin the specification stage by testing the linearity hypothesis.

3.1 Weak white noise and strong white noise testing

In view of (11), nonlinearities can be detected by testing if the linear innovations are iid or simply uncorrelated. We begin in Section 3.1.1 by testing if a sequence of observations (or the errors of a linear model) is a weak white noise or displays autocorrelation. In Section 3.1.2 we test if a weak white noise is strong or not.

3.1.1 Detection of autocorrelations

Assume that (ϵ_t) is a weak white noise $\mathrm{WN}(0, \sigma^2)$. The sample autocorrelations $\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$, where $\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-|h|} \epsilon_t \epsilon_{t+|h|}$, are expected to be close to $\rho(h) = \mathrm{Cov}(\epsilon_t, \epsilon_{t-h}) = 0$ for all $h \neq 0$. When a central limit theorem (CLT) applies (see Section 4 below for conditions ensuring the CLT), any vector of sample autocorrelations $\hat{\rho}_m = \{\hat{\rho}(1), \ldots, \hat{\rho}(m)\}'$ satisfies, as $n \to \infty$,

$$\sqrt{n}\hat{\rho}_m \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_m), \quad \Sigma_m(i,j) = \frac{1}{\sigma^4} \sum_{\ell=-\infty}^{\infty} E\epsilon_t \epsilon_{t+i} \epsilon_{t+\ell} \epsilon_{t+\ell+j}. \tag{12}$$

Note that for a strong white noise $\mathrm{IID}(0, \sigma^2)$, the asymptotic variance Σ_m is the identity matrix I_m. In this case the approximated 5% significance limits of the sample autocorrelations are $\pm 1.96/\sqrt{n}$ (as dotted line in Figure 4). These significance limits are also used extensively as a diagnostic check on the residuals of a fitted ARMA model. It is however important to note that these limits are not (asymptotically) valid when (ϵ_t) is only $\mathrm{WN}(0, \sigma^2)$, but not $\mathrm{IID}(0, \sigma^2)$.

Example 5 *Consider the ARCH(1) model* $\epsilon_t = \sigma_t \eta_t$, *with* η_t $\mathrm{IID}(0,1)$, $\kappa = E\eta_t^4$

and $\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2$. *Straightforward computations show that* $\Sigma_m(i,i) = 1 + (\kappa - 1)\alpha^i(1-\kappa\alpha^2)^{-1}$ *which may be quite different from the value* $\Sigma_m(i,i) = 1$ *obtained in the* $\mathrm{IID}(0,\sigma^2)$ *case.*

For checking the whiteness of a series, it is customary to test the assumption $H_0 : \rho(1) = \cdots = \rho(m) = 0$ by the so-called portmanteau tests, based on the Box and Pierce [25] statistic $Q_m^{BP} = n\sum_{i=1}^m \hat{\rho}^2(i)$ and on the Ljung-Box [88] statistic $Q_m^{LB} = n(n+2)\sum_{i=1}^m \hat{\rho}^2(i)/(n-i)$. Under the assumption that ϵ_t is $\mathrm{IID}(0,\sigma^2)$, the asymptotic distribution of these portmanteau statistics is χ_m^2, but this is not true for a general $\mathrm{WN}(0,\sigma^2)$. One can however work with a modified portmanteau statistic $Q_m = n\hat{\rho}_m'\hat{\Sigma}_m^{-1}\hat{\rho}_m$ which converges in law to a χ_m^2 whenever (12) holds and $\hat{\Sigma}_m$ is a consistent estimator of the nonsingular matrix Σ_m. The problem with the standard portmanteau tests based on Q_m^{LB} is that the white noise hypothesis can be rejected because the observations, though uncorrelated, are not independent (see Example 6 below).

The same difficulties hold when the portmanteau tests are applied to $\mathrm{ARMA}(p,q)$ residuals. The conventional $\chi^2_{m-(p+q)}$ distribution is no more valid in the presence of nonindependent innovations. In other words, the standard portmanteau goodness-of-fit tests are not reliable when the model is ARMA and the DGP is nonlinear. Is is however possible to modified these tests to take into account conditional heteroscedasticity or any other dependence in the linear innovations (see [89] and [59]).

Example 6 *The left graph of Figure 4 displays the autocorrelations of a simulation of length* $n = 5000$ *of an iid* $\mathcal{N}(0,1)$ *sequence. The right graph is analogous, but it concerns the GARCH(1,1) process*

$$\begin{cases} \epsilon_t = \sigma_t\eta_t, \quad \eta_t \sim \mathcal{N}(0,1) \\ \sigma_t^2 = 1 + 0.3\epsilon_{t-1}^2 + 0.55\sigma_{t-1}^2. \end{cases} \tag{13}$$

The thick dotted lines correspond to the true asymptotic 5% significance limits for the sample autocorrelations, whereas the fine horizontal dotted lines $\pm 1.96/\sqrt{n}$ *correspond to the asymptotic 5% significance limits for the sample autocorrelations of a strong white noise. In concrete applications the true significance limits are unknown, but can be estimated (see [105], [89] and [61]).*

Table 1 reports the results of the standard and modified portmanteau tests, for a simulation of length $n = 5000$ *of the GARCH(1,1) model (13). The standard portmanteau tests indicate that the strong white noise assumption must be rejected. One could think that this is due to the existence of non zero autocorrelations, and one could erroneously draw the conclusion that an ARMA model should be fitted to take into account these autocorrelations. The right conclusion is given by the modified portmanteau tests, which do not find strong evidence against the* $\mathrm{WN}(0,\sigma^2)$ *hypothesis.*

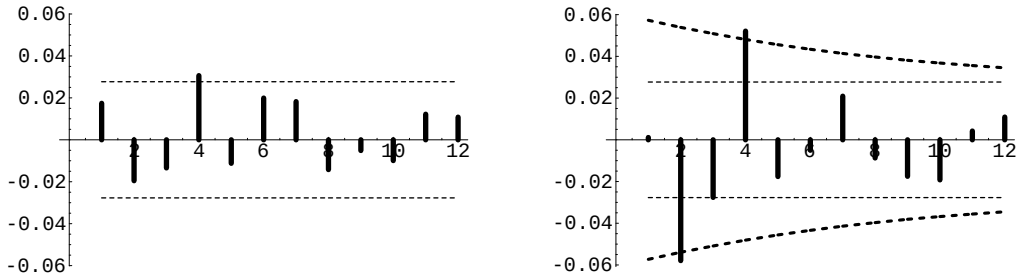


Figure 4. Autocorrelation of a strong white noise IID$(0, \sigma^2)$ (left graph) and of the weak white noise WN$((0, \sigma^2)$ defined by (13) (right graph).

Table 1
Portmanteau tests on a simulation of the GARCH(1,1) process (13).

Weak white noise tests based on Q_m						
m	1	2	3	4	5	6
$\hat{\rho}(m)$	0.00	-0.06	-0.03	0.05	-0.02	0.00
$\hat{\sigma}_{\hat{\rho}(m)}$	0.025	0.028	0.024	0.024	0.021	0.026
Q_m	0.00	4.20	5.49	10.19	10.90	10.94
$P(\chi^2_m > Q_m)$	0.9637	0.1227	0.1391	0.0374	0.0533	0.0902
Usual white noise tests						
m	1	2	3	4	5	6
$\hat{\rho}(m)$	0.00	-0.06	-0.03	0.05	-0.02	0.00
$\hat{\sigma}_{\hat{\rho}(m)}$	0.014	0.014	0.014	0.014	0.014	0.014
Q^{LB}_m	0.01	16.78	20.59	34.18	35.74	35.86
$P(\chi^2_m > Q^{LB}_m)$	0.9365	0.0002	0.0001	0.0000	0.0000	0.0000

3.1.2 Detection of serial dependence

We have seen that a process is linear if its innovation process (ϵ_t) is a strong IID$(0, \sigma^2)$ noise. It is well known that the two random variables ϵ_t and ϵ_{t-h} are independent if and only if $\text{Cor}\{\varphi(\epsilon_t), \vartheta(\epsilon_{t-h})\} = 0$ for all functions φ and ϑ such that $E\varphi^2(\epsilon_t) < \infty$ and $E\vartheta^2(\epsilon_t) < \infty$. It is thus natural to test the assumption that (ϵ_t) is a strong white noise by inspecting whether the squared process (ϵ_t^2), or other transformed processes of the form $\{\varphi(\epsilon_t)\}$, is correlated or not. McLeod and Li [91] showed that, when $\hat{\rho}_{\epsilon^2}(\cdot)$ is the sample autocorrelation function of the squared residuals $\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2$ of a strong ARMA model, the portmanteau statistics $Q^{LB}_{\epsilon^2,m} = n(n+2)\sum_{i=1}^m \hat{\rho}^2_{\epsilon^2}(i)$ is asymptotically χ^2_m distributed. Note that, contrary to Q^{LB}_m, the asymptotic distribution of $Q^{LB}_{\epsilon^2,m}$ does not depend on the number $p+q$ of estimated parameters. Numerous

other linearity tests are now available (see the book [83] or the paper [79] for recent references).

3.2 Testing linearity against a specific nonlinear model

Consider a nonlinear model which nests the linear model. Assume that the unknown parameter $\theta_0 = (\beta_0', \psi_0')'$ is such that the linearity hypothesis reduces to $H_0 : \psi_0 = 0$, with $\psi_0 \in \mathbb{R}^s$. Hypothesis testing is often based on the Wald, Score or Likelihood Ratio principle (see [52] and [65] for general references on these tests). The score test, also called Lagrange multiplier (ML) test or Rao test, is often very convenient because it does not require the estimation of the nonlinear model (which, in view of the identifiability problem already mentioned, is often difficult when H_0 holds true). This test only requires the estimation of the constrained estimator under H_0, denoted by $\hat{\theta}^c = (\hat{\beta}^{c'}, 0')'$. This estimator is often very simple, and sometimes coincides with an ordinary least-squares estimator of the form $\hat{\beta}^{c'} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}$.

3.2.1 General form of the LM test statistic

Under standard assumptions which will be discussed in Section 5, the conditional (quasi) log-likelihood takes the form

$$\ell_n(\theta) \overset{o_P(1)}{=} \sum_{t=1}^{n} \log f_\theta(Y_t \mid X_t, Y_{t-1}, \dots),$$

where $\theta = (\beta', \psi')'$ and $a \overset{c}{=} b$ stands for $a = b + c$. Moreover the score vector satisfies a CLT, and we have

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell_n(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left\{0, I := \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}\right\},$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left\{0, I^{-1} := \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}\right\}, \quad I^{22} = \left(I_{22} - I_{21}I_{11}^{-1}I_{12}\right)^{-1},$$

where $\hat{\theta}$ denotes the unconstrained quasi-maximum likelihood estimator (QMLE). Under H_0, a Taylor expansion yields

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell_n(\hat{\theta}) \overset{o_P(1)}{=} \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell_n(\hat{\theta}^c) + \frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\theta_0)\,\sqrt{n}(\hat{\theta} - \hat{\theta}^c). \tag{14}$$

The left-hand size is 0 when $\hat{\theta}$ is an interior point of the parameter space, and $\partial\ell_n(\hat{\theta}^c)/\partial\theta' = (0', \partial\ell_n(\hat{\theta}^c)/\partial\psi')$ for the same reason. Assume that

$n^{-1}\partial^2\ell_n(\theta_0)/\partial\theta\partial\theta'$ converges to $-I$. Then, under H_0, the first rows of (14) yield

$$\sqrt{n}(\hat{\beta}^c - \hat{\beta}) \overset{o_P(1)}{=} I_{11}^{-1} I_{12}\sqrt{n}\hat{\psi} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I_{11}^{-1} I_{12} I^{22} I_{21} I_{11}^{-1}\right)$$

and the last s rows of (14) yield

$$\begin{aligned}\Delta_n^c &:= \frac{1}{\sqrt{n}}\frac{\partial}{\partial\psi}\ell_n(\hat{\theta}^c) \overset{o_P(1)}{=} I_{21}\sqrt{n}(\hat{\beta} - \hat{\beta}^c) + I_{22}\sqrt{n}\hat{\psi} \\ &\overset{o_P(1)}{=} (-I_{21}I_{11}^{-1}I_{12} + I_{22})\sqrt{n}\hat{\psi} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_\Delta)\end{aligned}$$

where $\Sigma_\Delta = (I^{22})^{-1}$.

A Rao score-type statistic is then given by $\mathrm{LM}_n = (\Delta_n^c)'\,\hat{\Sigma}_\Delta^{-1}\Delta_n^c$ where $\hat{\Sigma}_\Delta$ denotes any H_0-consistent estimator of Σ_Δ. This statistic follows asymptotically a χ_s^2 distribution under the null. This leads to critical regions of the form $\{\mathrm{LM}_n > \chi_s^2(1-\alpha)\}$.

3.2.2 LM statistic when $I_{21} = 0$

The LM test can sometimes be carried out very easily. Consider a general conditionally homoscedastic nonlinear model of the form $Y_t = F_\theta(W_t) + \epsilon_t$, where $W_t = (X_t, Y_{t-1}, \dots)$ depends on an exogenous vector X_t and on the past values of the endogenous variable, and ϵ_t is $\mathrm{IID}(0, \sigma^2)$, so that W_t and ϵ_t are independent. We assume that $\sigma^2 = \sigma^2(\beta)$ does not depend on ψ. With a Gaussian quasi-likelihood, we have

$$\Delta_n^c = \frac{1}{\sqrt{n}\hat{\sigma}^{c\,2}}\sum_{t=1}^n \epsilon_t(\hat{\theta}^c)\frac{\partial}{\partial\psi}F_{\hat{\theta}^c}(W_t) = \frac{1}{\sqrt{n}\hat{\sigma}^{c\,2}}\mathbf{F}_\psi'\hat{\mathbf{U}}^c$$

with $\epsilon_t(\theta) = Y_t - F_\theta(W_t)$, $\hat{\sigma}^{c\,2} = \sigma^2(\hat{\beta}^c)$, $\hat{\epsilon}_t^c = \epsilon_t(\hat{\theta}^c)$ and $\hat{\mathbf{U}}^c = (\hat{\epsilon}_1^c, \dots, \hat{\epsilon}_n^c)'$. Under the assumption that the information matrix I is block-diagonal, *i.e.* when $I_{12} = \sigma^{-2}E\left\{\partial F_{\theta_0}(W_t)/\partial\beta\right\}\left\{\partial F_{\theta_0}(W_t)/\partial\psi'\right\} = 0$, the asymptotic distribution of Δ_n^c is equal to that of $n^{-1/2}\sigma^{-2}\sum_{t=1}^n \epsilon_t\partial F_{\theta_0}(W_t)/\partial\psi$, and one can take $\hat{\Sigma}_\Delta = n^{-1}\left(\hat{\sigma}^c\right)^{-4}\hat{\mathbf{U}}^{c\prime}\hat{\mathbf{U}}^c\left(n^{-1}\mathbf{F}_\psi'\mathbf{F}_\psi\right) \overset{o_P(1)}{=} \mathbf{F}_\psi'\mathbf{F}_\psi/(\hat{\mathbf{U}}^{c\prime}\hat{\mathbf{U}}^c)$ as a consistent estimator of Σ_Δ. We then obtain the following simple version of the score statistic

$$\mathrm{LM}_n = n\frac{\hat{\mathbf{U}}^{c\prime}\mathbf{F}_\psi\left(\mathbf{F}_\psi'\mathbf{F}_\psi\right)^{-1}\mathbf{F}_\psi'\hat{\mathbf{U}}^c}{\hat{\mathbf{U}}^{c\prime}\hat{\mathbf{U}}^c}, \tag{15}$$

which is n times the uncentered coefficient of determination of the regression of $\hat{\epsilon}_t^c$ on the variables $\partial F_{\hat{\theta}^c}(W_t)/\partial\psi_i$ for $i = 1, \dots, s$.

3.2.3 LM test with auxiliary regressions

When I_{12} is not assumed to be zero, and when σ^2 is a nuisance parameter which does not depend on the parameter of interest $\theta = (\beta', \psi')'$, one can estimate Σ_Δ by

$$\hat{\Sigma}^*_\Delta = \frac{1}{n\hat{\sigma}^{c\,4}} \hat{\mathbf{U}}^{c\prime} \hat{\mathbf{U}}^c \left(n^{-1} \mathbf{F}'_\psi \mathbf{F}_\psi - n^{-1} \mathbf{F}'_\psi \mathbf{F}_\beta \left(n^{-1} \mathbf{F}'_\beta \mathbf{F}_\beta \right)^{-1} n^{-1} \mathbf{F}'_\beta \mathbf{F}_\psi \right),$$

where

$$\mathbf{F}'_\beta = \left(\frac{\partial F_{\hat{\theta}^c}(W_1)}{\partial \beta} \cdots \frac{\partial F_{\hat{\theta}^c}(W_n)}{\partial \beta} \right).$$

Because the initial model is linear under the constraint H_0, we generally have $\hat{\mathbf{U}}^c = \mathbf{Y} - \mathbf{F}_\beta \hat{\beta}^c$ and $\hat{\sigma}^{c\,2} = \hat{\mathbf{U}}^{c\prime} \hat{\mathbf{U}}^c / n$ with $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and $\hat{\beta}^c = \left(\mathbf{F}'_\beta \mathbf{F}_\beta \right)^{-1} \mathbf{F}'_\beta \mathbf{Y}$, up to negligible terms (unknown initial values entail that the conditional QMLE does not exactly coincide with the least squares estimator (LSE) under the null).

Now consider the auxiliary linear regression

$$\mathbf{Y} = \mathbf{F}_\beta \beta^* + \mathbf{F}_\psi \psi^* + \mathbf{U}. \tag{16}$$

In this auxiliary regression, the LM test statistic of the hypothesis $H_0^* : \psi^* = 0$ is equal to

$$\begin{aligned} \mathrm{LM}^*_n &= n^{-1} \left(\hat{\sigma}^c \right)^{-4} \hat{\mathbf{U}}^{c\prime} \mathbf{F}_\psi \left(\hat{\Sigma}^*_\Delta \right)^{-1} \mathbf{F}'_\psi \hat{\mathbf{U}}^c \\ &= \left(\hat{\sigma}^c \right)^{-2} \hat{\mathbf{U}}^{c\prime} \mathbf{F}_\psi \left(\mathbf{F}'_\psi \mathbf{F}_\psi - \mathbf{F}'_\psi \mathbf{F}_\beta \left(\mathbf{F}'_\beta \mathbf{F}_\beta \right)^{-1} \mathbf{F}'_\beta \mathbf{F}_\psi \right)^{-1} \mathbf{F}'_\psi \hat{\mathbf{U}}^c, \end{aligned}$$

(see [52] eq. (24)), which is precisely the LM test statistic of the hypothesis $H_0 : \psi = 0$ in the initial model. It is well known (see [52] eq. (27)) that the LM statistic which is associated with $H_0^* : \psi^* = 0$ in (16) can also be written as

$$\mathrm{LM}^*_n = n \frac{\hat{\mathbf{U}}^{c\prime} \hat{\mathbf{U}}^c - \hat{\mathbf{U}}' \hat{\mathbf{U}}}{\hat{\mathbf{U}}^{c\prime} \hat{\mathbf{U}}^c}, \tag{17}$$

where $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{F}_\beta \hat{\beta}^* - \mathbf{F}_\psi \hat{\psi}^* =: \mathbf{Y} - \mathbf{F} \hat{\theta}^*$, with $\hat{\theta}^* = \left(\mathbf{F}'\mathbf{F} \right)^{-1} \mathbf{F}'\mathbf{Y}$. We finally obtain the Breusch-Godfrey form of the LM statistic by interpreting LM^*_n in (17) as n times the coefficient of determination of the new auxiliary linear regression

$$\hat{\mathbf{U}}^c = \mathbf{F}_\beta \gamma + \mathbf{F}_\psi \psi^{**} + \mathbf{V}, \tag{18}$$

where $\hat{\mathbf{U}}^c$ is the vector of the residuals of the regression of $\mathbf{Y}$ on the columns of $\mathbf{F}_\beta$. Indeed, in the two regressions (16) and (18), the vector of the residuals is $\hat{\mathbf{V}} = \hat{\mathbf{U}}$, because $\hat{\beta}^* = \hat{\beta}^c + \hat{\gamma}$ and $\hat{\psi}^* = \hat{\psi}^{**}$. Finally note that the coefficient of determination is centered (the standard R^2 provided by most of the packaged computer programs) when a column of $\mathbf{F}_\beta$ is constant.

Example 7 *Consider a stationary process (Y_t) satisfying the bilinear model (5) where $(\epsilon_t) \sim \text{IID}(0, \sigma^2)$, and ϵ_t is independent of Y_{t-i} for $i > 0$. The linearity assumption writes $H_0 : \psi = 0$, with $\psi = (c_{11}, \ldots, c_{PQ})'$. Note that, with obvious notations,*

$$\frac{\partial F_{\theta^c}(W_t)}{\partial \beta} = \left(1 + \sum_{j=1}^{q} b_j B^j\right)^{-1} \{Y_{t-1}, \ldots, Y_{t-p}, \epsilon_{t-1}(\theta^c), \ldots, \epsilon_{t-q}(\theta^c)\}'$$

and

$$\frac{\partial F_{\theta^c}(W_t)}{\partial \psi} = \left(1 + \sum_{j=1}^{q} b_j B^j\right)^{-1} \{Y_{t-1}\epsilon_{t-1}(\theta^c), \ldots, Y_{t-P}\epsilon_{t-Q}(\theta^c)\}'.$$

If we assume $q = 0$ and the symmetry condition $E\epsilon_t^3 = 0$ we have $I_{12} = 0$, and (15) applies. Thus, in the case $q = 0$ and $E\epsilon_t^3 = 0$ the LM test can be carried out as follows: 1) fit an AR(p) model and compute the residuals $\hat{\epsilon}_t^c$, 2) regress $\hat{\epsilon}_t^c$ on $Y_{t-1}\hat{\epsilon}_{t-1}^c, \ldots, Y_{t-P}\hat{\epsilon}_{t-Q}^c$ and compute the uncentered R^2 of this regression. We then reject H_0 when $\text{LM}_n = nR^2 > \chi^2_{PQ}(1-\alpha)$.

If we no longer assume $E\epsilon_t^3 = 0$ but we still assume $q = 0$ (as in [120] or [113]), then (17) holds and the LM test can be implemented as follows: 1') fit an AR(p) model, compute the residuals $\hat{\epsilon}_t^c$ and the residual sum of squares RSS^c, 2') regress $\hat{\epsilon}_t^c$ on $Y_{t-1}, \ldots, Y_{t-p}$ and $Y_{t-1}\hat{\epsilon}_{t-1}^c, \ldots, Y_{t-P}\hat{\epsilon}_{t-Q}^c$, and compute the residual sum of squares RSS. We then reject H_0 when

$$\text{LM}_n = n(\text{RSS}^c - \text{RSS})/\text{RSS}^c > \chi^2_{PQ}(1-\alpha).$$

The F test is an alternative which is asymptotically equivalent but might perform better in finite sample. With this test we reject H_0 when

$$\text{F}_n = (n - p - PQ)PQ^{-1}(\text{RSS}^c - \text{RSS})/\text{RSS}$$

is greater than the $1-\alpha$ quantile of the Fisher-Snedecor $\mathcal{F}(PQ, n-p-PQ)$ distribution.

3.3 Testing linearity when the model is not identified under the null

Many nonlinear models contain nuisance parameters which are not identified under the null assumption of linearity. As an illustrative example, consider a SETAR model of the form

$$Y_t = \nu^{(1)} + a_1^{(1)} Y_{t-1} + \left(\nu^{(2)} + a_1^{(2)} Y_{t-1}\right) \mathbf{1}_{\{Y_{t-1} > r\}} + \epsilon_t, \tag{19}$$

where (ϵ_t) is $\text{IID}(0, \sigma^2)$. The null assumption of interest is that Y_t is a stationary strong AR(1) process. This assumption can be written as

$$H_0 : \ \nu^{(2)} = a_1^{(2)} = 0.$$

Under H_0 the threshold parameter r does not exist. As a consequence the usual estimators, in particular the quasi-maximum likelihood (QML) and least squares (LS) estimators, have nonstandard behaviors. In particular any reasonable estimator $\hat{r}$ of r should be consistent under the alternative H_1 of a SETAR model, but should not converge to any value under H_0. The usual tests, such as the LM test defined in Section 3.2, are also affected by the lack of identification of the parameter r under the null. A coarse solution is obtained by fixing an arbitrary value for the threshold r. We then work with the pointwise LM statistic

$$\mathrm{LM}_n(r) = n\frac{\hat{\sigma}^{c\,2} - \hat{\sigma}_r^2}{\hat{\sigma}^{c\,2}},$$

where $\hat{\sigma}^{c\,2}$ is the mean of the squares of the residuals of the AR(1) model implied by the null, and $\hat{\sigma}_r^2$ is the residual mean square of the SETAR model (19) with given threshold r. One can also employ pointwise Wald or likelihood ratio (LR) statistics of the form

$$\mathrm{W}_n(r) = n\frac{\hat{\sigma}^{c\,2} - \hat{\sigma}_r^2}{\hat{\sigma}_r^2} \qquad \text{and} \qquad \mathrm{LR}_n(r) = n\log\frac{\hat{\sigma}^{c\,2}}{\hat{\sigma}_r^2}.$$

Under regularity conditions similar to those discussed in Section 5 below, all these statistics are asymptotically χ_2^2 distributed under H_0. The resulting tests are often consistent, even for alternatives such that the true threshold is not equal to the chosen value of r. It should however be underlined that the choice of r is unpleasant, because no a priori reasonable value is generally available. Moreover the choice of r has an obvious impact on the power of the tests (the power is likely to be low for alternative models in which the actual threshold is far from r). The threshold can be estimated by least squares as

$$\hat{r} = \arg\min_{r\in[\underline{r},\overline{r}]} \hat{\sigma}_r^2,$$

where $\underline{r}$ and $\overline{r}$ are given constants such that $\underline{r} < \overline{r}$. In [74], $\underline{r}$ and $\overline{r}$ are chosen to be the 15th and 85th percentiles of the empirical distribution of the observations. The standard Wald, LM and LR test statistics then satisfy

$$\mathrm{W}_n = \mathrm{W}_n(\hat{r}), \qquad \mathrm{LM}_n = \mathrm{LM}_n(\hat{r}), \qquad \mathrm{LR}_n = \mathrm{LR}_n(\hat{r}).$$

Figure 5 shows that, under the null the distribution of W_n is completely different from that of a χ_2^2. This is a consequence of the nonstandard behaviour of $\hat{r}$ under H_0. Following [45], [46] and [74] the common asymptotic distribution of the 3 statistics W_n, LM_n and LR_n is a functional of a continuous-time gaussian process under H_0.

Since $\mathrm{W}_n = \sup_{r\in[\underline{r},\overline{r}]} \mathrm{W}_n(r)$, the standard Wald statistic can be viewed as a supremum test statistic. The same interpretation holds for the LM and LR statistics. Other functionals of the pointwise statistics are suggested in [9].

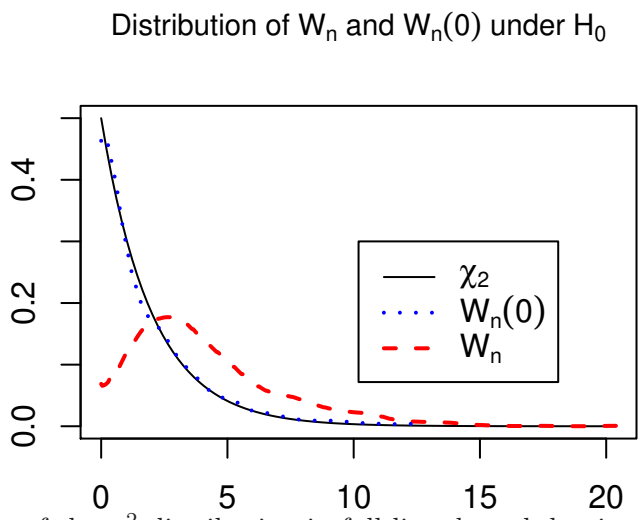


Figure 5. Density of the χ^2_2 distribution in full line, kernel density estimator of the distribution of $\mathrm{W}_n(0)$ in dotted line and of W_n in dashed line. The two density estimators are obtained by computing the statistics on $N = 1,000$ independent replications of simulations of length $n = 200$ of an iid $\mathcal{N}(0,1)$ sequence.

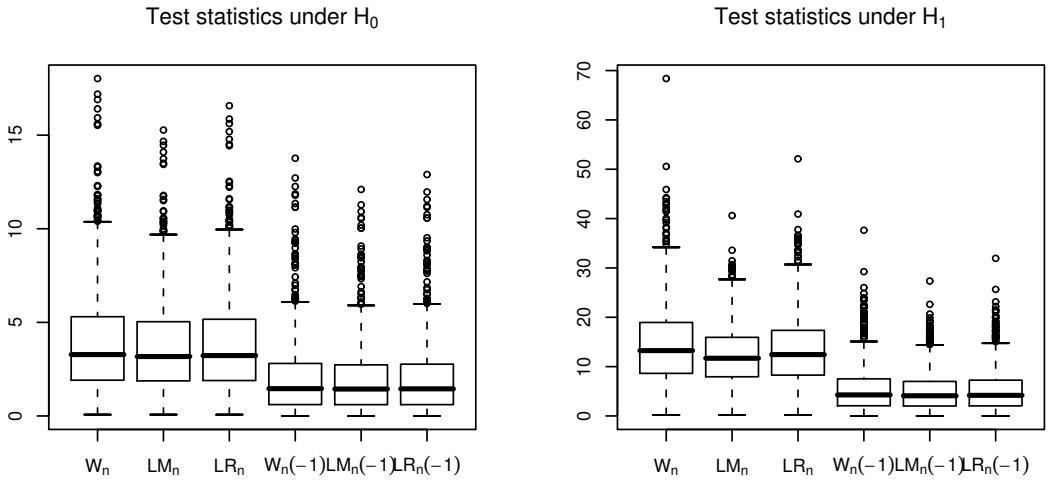


Figure 6. Distribution of some test statistics under H_0 ($N = 1,000$ independent simulations of length $n = 100$ of the AR(1) model $Y_t = 0.7Y_{t-1} + \epsilon_t$ where (ϵ_t) is IID $\mathcal{N}(0,1)$) and H_1 ($N = 1,000$ independent simulations of length $n = 100$ of the SETAR model (19) with $(\nu^{(1)}, a_1^{(1)}, \nu^{(2)}, a_1^{(2)}, r) = (0, 0.9, -2, -0.7, -2)$).

The test proposed by [90] (hereafter LST) is the most commonly used for testing linearity against smooth transition autoregressive models. It also makes sense to use the LST test for testing linearity against SETAR models. For instance, the model (19) can be approximated by a logistic smooth transition

Table 2
Relative frequency of rejection of the linearity hypothesis H_0 for tests with nominal level $\alpha = 5\%$, based on $N = 1,000$ independent replications of simulations of length $n = 100$.

Design		Wald, score and Likelihood Ratio					LST tests	
	tests	$r = \hat{r}$	$r = 1$	$r = 0$	$r = -1$	$r = -2$	1-LST	3-LST
	$W_n(r)$	5.6	5.7	4.6	4.7	4.4	3.1	4.0
H_0	$LM_n(r)$	3.3	4.1	3.7	3.4	3.9	2.8	3.0
	$LR_n(r)$	4.7	4.8	4.4	4.0	4.1	3.0	3.5
	$W_n(r)$	70.3	0.3	6.1	34.7	82.0	65.0	61.2
H_1	$LM_n(r)$	64.8	0.3	5.4	31.3	79.1	64.5	59.7
	$LR_n(r)$	67.8	0.3	5.9	33.0	80.3	64.7	60.6

Design H_0: AR(1) model $Y_t = 0.9Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$

Design H_1: SETAR model (19) with $r = -2$, as defined in Figure 6

autoregressive model

$$Y_t = \nu^{(1)} + a_1^{(1)} Y_{t-1} + \left(\nu^{(2)} + a_1^{(2)} Y_{t-1}\right) G(\gamma, r, Y_{t-1}) + \epsilon_t,$$

when, in the logistic function $G(\gamma, r, x) = [1 + \exp\{-\gamma(x - r)\}]^{-1}$, the slope parameter γ is large. In its simplest version, the LST test consists in testing $\phi = 0$ in the auxiliary model

$$Y_t = \nu^{(1)} + a_1^{(1)} Y_{t-1} + \phi Y_{t-1}^2 + \epsilon_t. \tag{20}$$

The auxiliary model is obtained by using the Taylor expansion $G(\gamma, r, x) = 1/2 + (x - r)\gamma/4 + o(\gamma)$ and a reparameterization of the model (see [90]). Using a third-order Taylor expansion, instead of a first-order one, the LST approach leads to test $\phi_1 = \phi_2 = \phi_3 = 0$ in the auxiliary model

$$Y_t = \nu^{(1)} + a_1^{(1)} Y_{t-1} + \phi_1 Y_{t-1}^2 + \phi_2 Y_{t-1}^3 + \phi_3 Y_{t-1}^4 + \epsilon_t. \tag{21}$$

The test based on the auxiliary model (20) is denoted by 1-LST, and the one based on (21) is denoted by 3-LST. At the asymptotic level α, the critical values of these tests are respectively the quantiles $\chi_1^2(1 - \alpha)$ and $\chi_3^2(1 - \alpha)$. Note that the 2-LST version does not exist because, around 0, the second-order Taylor expansion of $G(\cdot, r, x)$ coincides with the first-order one.

Figure 6 and Table 2 summarize the results of simulation experiments which compare the behavior of the different tests. In Table 2, the critical values of the supremum tests have been determined by means of simulations. Figure 6 shows that, as expected, for any fixed value r the 3 pointwise statistics

$W_n(r)$, $LM_n(r)$ and $LR_n(r)$ have similar behaviors under the null and local alternatives, but may be quite different for alternatives which are far from the null. The same remark holds for the supremum statistics W_n, LM_n and LR_n. The behavior of the statistics based on the data dependent value $r = \hat{r}$ is completely different from that of the pointwise statistics based on a data independent value of r. Table 2 shows that the supremum tests are much more powerful than the pointwise tests when the latter are based on a value r which is far from the true alternative ($r = -2$ for the displayed experiments), but they are of course less powerful than the pointwise tests based on the true value $r = -2$. We also note that, although the LST tests are extremely simple and easy to implement, their performance is comparable with that of the supremum tests, at least on the set of Monte Carlo experiments we considered.

4 Probabilistic tools

In this section we present some probabilistic tools which may be useful in nonlinear model analysis.

4.1 A strict stationarity condition

In [22] (see also [26] and [23]), Bougerol and Picard derived a necessary and sufficient condition for the existence of a strictly stationary solution to the linear stochastic recurrent equation

$$Z_t = A_t Z_{t-1} + B_t, \quad t \in \mathbb{Z}, \tag{22}$$

where A_t is a $d \times d$ random matrix, B_t is a random vector, and $(A_t, B_t)_{t\in\mathbb{Z}}$ is an iid sequence. Under mild assumptions, the exists a nonanticipative stationary solution to (22) if and only if

$$\boldsymbol{\gamma} := \inf_{n\in\mathbb{N}^*} \frac{1}{n} E\left(\log \|A_n A_{n-1} \cdots A_1\|\right) \overset{a.s.}{=} \lim_{n\to\infty} \frac{1}{n} \log \left\| \prod_{i=1}^{n} A_{t-i} \right\| < 0. \tag{23}$$

Under this condition the solution is ergodic, which means that the law of large numbers applies: as $n \to \infty$

$$\frac{1}{n} \sum_{t=1}^{n} \varphi(\dots, Z_{t-1}, Z_t, Z_{t+1}, \dots) \to E\varphi(\dots, Z_{t-1}, Z_t, Z_{t+1}, \dots) \tag{24}$$

provided the last expectation exists (see [20]). The coefficient $\boldsymbol{\gamma}$ is called the top Lyapounov exponent of the sequence (A_t), and can be evaluated by simulations (see [40] and [41]). The condition (23) is sufficient when $(A_t, B_t)_{t\in\mathbb{Z}}$ is

strictly stationary and ergodic. This condition is directly applicable for processes (Y_t) having a state-space representation of the form $Y_t = HZ_t$ with (22).

Example 8 *Consider a nonlinear model of the form*

$$Y_t = \epsilon_t + c_1 Y_{t-1}\epsilon_t + c_2 Y_{t-2}\epsilon_t^2, \quad (\epsilon_t) \sim \text{IID}(0, \sigma^2). \tag{25}$$

We have (22) with

$$Z_t = \begin{pmatrix} Y_t \\ Y_{t-1} \end{pmatrix}, \quad A_t = \begin{pmatrix} c_1\epsilon_t & c_2\epsilon_t^2 \\ 1 & 0 \end{pmatrix}, \quad B_t = \begin{pmatrix} \epsilon_t \\ 0 \end{pmatrix}.$$

We deduce that (25) admits a nonanticipative strictly stationary solution if and only if the top-Lyapounov exponent $\boldsymbol{\gamma}$ of (A_t) is strictly negative. Figure 7 displays an estimation of the strict stationarity region, obtained by evaluating $\boldsymbol{\gamma}$ from simulations of (A_t). The strict stationary curve passes at the points $(c_1, c_2) = (\pm e^{-E\log|\epsilon_t|}, 0)$ and $(c_1, c_2) = (0, \pm e^{-E\log\epsilon_t^2})$. Other computations show that, when ϵ_t is gaussian, the second order stationarity region is given by the constraint $c_1^2 E\epsilon_t^2 + c_2^2 E\epsilon_t^4 < 1$.

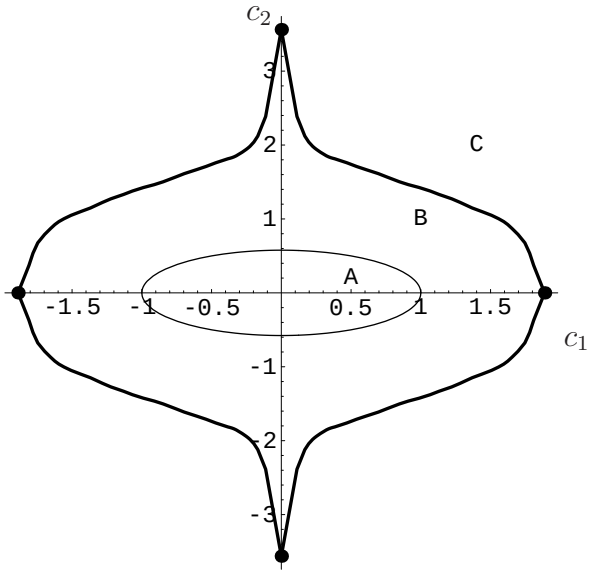


Figure 7. Strict and second-order stationarity regions of the model (25) with $\epsilon_t \sim \mathcal{N}(0,1)$. A: second order stationarity, A ∪ B: strict stationarity, B ∪ C: non second-order stationarity, and C: non strict stationary.

4.2 Second order stationarity and existence of moments

Under the condition $\boldsymbol{\gamma} < 0$, the strict stationary solution to (22) writes

$$Z_t = B_t + \sum_{k=1}^{\infty} A_t \cdots A_{t-k+1} B_{t-k} \quad a.s.$$

From the Cauchy criterion, the vector is well-defined in L^2 if $\|A_t \cdots A_{t-k+1} B_{t-k}\|_2$ exists and converges to 0 at an exponential rate as $k \to \infty$. Using the iid assumption made on (A_t, B_t) and elementary matrix manipulations, we have

$$\begin{aligned} E\|A_t \cdots A_{t-k+1} B_{t-k}\|^2 &= E\left(B'_{t-k} A'_{t-k+1} \cdots A'_t A_t \cdots A_{t-k+1} B_{t-k}\right) \\ &= E\left\{B'_{t-k} \otimes B'_{t-k}\right\} \left\{E\left(A'_t \otimes A'_t\right)\right\}^k \operatorname{vec} \mathrm{I}, \end{aligned}$$

where I denotes the identity matrix of size equal to the dimension of Z_t. Denoting by $\rho(M)$ the spectral radius of a square matrix M, we deduce that, provided $E\|B_t\|^2 < \infty$,

$$\rho\left\{E\left(A_t \otimes A_t\right)\right\} < 1 \tag{26}$$

is a sufficient condition for the existence of a second order stationary solution. In the previous argument, we check that (26) entails (23), using the Jensen inequality. The same elementary technique can be used to obtain conditions for the existence of higher order moments, and can sometimes be adapted in cases where (A_t, B_t) is not iid (see *e.g.* [60]).

Example 9 *Consider the simple Markov-switching model*

$$Y_t = \epsilon_t + a(1) Y_{t-1} \mathbf{1}_{\Delta_t = 1} + a(2) Y_{t-1} \mathbf{1}_{\Delta_t = 2}, \tag{27}$$

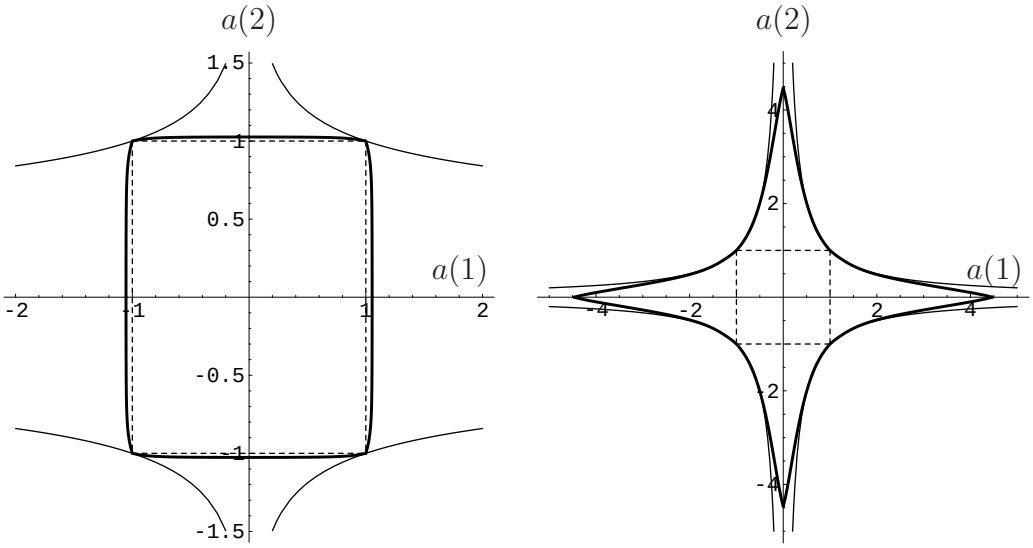
where ϵ_t is a strong white noise and Δ_t is an irreducible and aperiodic stationary Markov chain with state-space $\{1, 2\}$, transition probabilities $p(i, j) = P(\Delta_t = j \mid \Delta_{t-1} = i)$ and stationary probabilities $\pi(i) = P(\Delta_t = i)$. It is easy to check that $\sum_{i=1}^2 \pi(i) \log |a(i)| < 0$ is a sufficient condition for strict stationary. This condition can be interpreted as an average of the stationarity constraint over the 2 regimes, and involves the transition probabilities $p(i, j)$ only through the stationary probabilities $\pi(i)$. The (necessary and sufficient) second order stationarity condition (see [60]) is

$$\rho(A) < 1, \quad A = \begin{pmatrix} p(1,1)a^2(1) & p(2,1)a^2(1) \\ p(1,2)a^2(2) & p(2,2)a^2(2) \end{pmatrix}.$$

*Figure 8 displays these stationarity regions. It can be checked that $\rho(A) < 1$ is equivalent to (26) when $p(i, j) = \pi(j)$ for all j (*i.e. *when (Δ_t) is iid).*

4.3 Mixing coefficients

For statistical inference, strict stationarity and ergodicity are generally not sufficient assumptions, and it may be useful to know if a given process possesses mixing properties. Mixing is one way to characterize the decrease of dependence when the variables become sufficiently far apart (see *e.g.* [44]).



$p(1,1) = 0.8$ and $p(2,2) = 0.95$. $p(1,1) = p(2,2) = 0.05$.

Figure 8. Stationarity regions of the Markov-switching model (27). The second-order stationarity region is the bounded region containing the square $[-1,1] \times [-1,1]$ (displayed as dotted line). The unbounded region delimited by the 4 curves corresponds to strict stationary models.

More precisely the (strong) α-mixing coefficients of a process (Z_t) are defined by

$$\alpha_Z(k) = \sup_t \sup_{A \in \sigma(Z_s, s \leq t),\ B \in \sigma(Z_s, s \geq t+k)} |P(A \cap B) - P(A)P(B)|, \tag{28}$$

where $\sigma(Z_s, s \leq t)$ denotes the information set generated by the past at the time t, and $\sigma(Z_s, s \geq t+k)$ denotes the information set generated by the future at the time $t+k$. There exist other mixing coefficients, in particular the β-mixing coefficients which are defined by

$$\beta_Z(k) = \sup_t \ E\left\{ \sup_{B \in \sigma(Z_s, s \geq t+k)} |P\{B \mid \sigma(Z_s, s \leq t)\} - P(B)| \right\}. \tag{29}$$

When (Z_t) is stationary, the term $\sup_t$ can be omitted in the definitions (28) and (29). The process is said to be α-mixing (resp. β-mixing) if $\lim_{k\to\infty} \alpha_Z(k) = 0$ (resp. $\lim_{k\to\infty} \beta_Z(k) = 0$). We have $\alpha_Z(k) \leq \beta_Z(k)$, so that β-mixing implies α-mixing. If $Y = (Y_t)$ is a process such that $Y_t = f(Z_t, \ldots, Z_{t-r})$ for some measurable function f and some integer $r \geq 0$, then $\sigma(Y_t,\ t \leq 0) \subset \sigma(Z_t,\ t \leq 0)$ and $\sigma(Y_t,\ t \geq s) \subset \sigma(Z_t,\ t \geq s - r)$. Thus

$$\alpha_Y(k) \leq \alpha_Z(k-r) \quad \text{and} \quad \beta_Y(k) \leq \beta_Z(k-r) \quad \text{for all } k \geq r.$$

The α-mixing coefficient between two σ-fields $\mathcal{A}$ and $\mathcal{B}$ is defined by

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A},\ B \in \mathcal{B}} |P(A \cap B) - P(A)P(B)|.$$

Let p, q and r be 3 positive numbers such that $p^{-1} + q^{-1} + r^{-1} = 1$. Davydov [47] showed the following inequality

$$|\mathrm{Cov}(U, V)| \leq K_0 \|U\|_p \|V\|_q \left[\alpha \left\{\sigma(U), \sigma(V)\right\}\right]^{1/r}, \tag{30}$$

where $\|U\|_p^p = EU^p$ and K_0 is an universal constant. Davydov [47] proposed $K_0 = 12$. Rio [106] obtained a sharper inequality involving the quantile functions of U and V, and showed that one can take $K_0 = 4$ in (30). Note that (30) entails that the autocovariance function $\Gamma_Z(h) \to 0$ as $|h| \to \infty$, when Z is a stationary α-mixing process (with moments of order greater than 2).

In statistical applications, the α-mixing assumption is convenient because it implies a central limit theorem. Herrndorf [77] showed that that under the assumptions $EZ_t = 0$ and

$$\sup_t \|Z_t\|_{2+\nu} < \infty, \qquad \sum_{h=0}^{\infty} \{\alpha_Z(h)\}^{\nu/(2+\nu)} < \infty \quad \text{for some } \nu > 0,$$

we have

$$n^{-1/2} \sum_{t=1}^{n} Z_t \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2), \quad \text{when} \quad \sigma^2 = \lim_{n\to\infty} \mathrm{Var}\left(n^{-1/2} \sum_{t=1}^{n} Z_t\right) > 0.$$

4.4 Geometric ergodicity and mixing properties

A way to check mixing properties, or to find out stationarity conditions, is to use the Markov chain theory (see the papers by [116] and [56], and the book by [95]).

Consider a Markov chain $(Z_t)_{t\in\mathbb{N}}$ with state space $(E, \mathcal{E})$, where E is a subset of $\mathbb{R}^d$ and $\mathcal{E}$ is a Borel σ-field on E. Let $P^t(x, B) = P(Z_t \in B \mid Z_0 = x)$ be the $t-$step transition probability of moving from $x \in E$ to the set $B \in \mathcal{E}$ in t steps. The Markov chain (Z_t) is said to be geometrically ergodic if there exist $\rho \in (0, 1)$ and a measure π such that

$$\forall x \in E, \quad \rho^{-t} \|P^t(x, \cdot) - \pi(\cdot)\| \to 0 \qquad \text{as } t \to \infty,$$

where $\|\cdot\|$ is the total variation norm. A consequence of geometric ergodicity is β-mixing, and hence α-mixing, with geometric rate. Moreover $(Z_t)_{t\in\mathbb{N}}$ is stationary when the distribution of Z_0 is the invariant probability π.

Let ϕ be a non trivial σ-finite measure on $(E, \mathcal{E})$. The chain (Z_t) is ϕ-irreducible if for all $x \in E$ and all $B \in \mathcal{E}$ such that $\phi(B) > 0$ there exists $t \in \{1, 2, \dots\}$ such that $P^t(x, B) > 0$. We say that (Z_t) is a Feller Markov chain when

the function $x \mapsto E\{g(Z_t)|Z_{t-1} = x\}$ is continuous for every bounded and continuous function g on E.

Feigin and Tweedie ([56], Theorem 1) showed that (Z_t) is geometrically ergodic when (i) (Z_t) is a Feller Markov chain, (ii) (Z_t) is ϕ-irreducible for some non trivial σ-finite measure ϕ on $(E, \mathcal{E})$, (iii) there exists a compact set $C \subset E$ such that $\phi(C) > 0$ and a non-negative continuous function $V : E \to \mathbb{R}$ such that

$$V(x) \geq 1, \qquad \forall x \in C \tag{31}$$

and for some $c > 0$

$$E\{V(Z_t)|Z_{t-1} = x\} \leq (1-c)V(x), \qquad \forall x \notin C. \tag{32}$$

As a consequence of Theorem 2 in Feigin and Tweedie ([56]), we have $E_\pi V(Z_t) < \infty$ when, in addition to (i)-(iii), the test function V satisfies

$$\text{iv)} \ \sup E\{V(Z_t)|Z_{t-1} = x\} < \infty, \quad x \in C. \tag{33}$$

Example 10 *Consider an EXPAR(1) model of the form*

$$Y_t = \left\{a + b\exp(-\gamma Y_{t-1}^2)\right\} Y_{t-1} + \epsilon_t \quad (\epsilon_t) \sim \text{IID}(0, \sigma^2), \tag{34}$$

where $\gamma > 0$ and $|a| < 1$. It is clear that $(Y_t)_{t\in\mathbb{N}}$ is a Markov chain with state space $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. Assume that ϵ_t has a strictly positive density over $\mathbb{R}$ and that $E|\epsilon_t| < \infty$. We will show that (Y_t) is geometrically ergodic, by checking (i)-(iii) above. The result being well known for $b = 0$, we assume $b \neq 0$. We have $E\{g(Y_t) \mid Y_{t-1} = y\} = E\ g\left\{\epsilon_t + \left(a + be^{-\gamma y^2}\right) y\right\}$. The dominated convergence theorem then shows that (Y_t) is a Feller chain, so that (i) is checked. Given $Y_0 = y$, the law of $Y_1 = \epsilon_1 + \left(a + be^{-\gamma y^2}\right) y$ admits a strictly positive density with respect to the Lebesgue measure λ. Thus (Y_t) is λ-irreducible and (ii) is checked. Let $V(y) = 1 + |y|$, a constant c such that

$$0 < c < \min\left\{\frac{1-|a|}{2}, |b|\right\}$$

and the compact interval

$$C = \left\{y : \ y^2 \leq \max\left\{\frac{-1}{\gamma}\log\left(\frac{c}{|b|}\right), \left(\frac{E|\epsilon_t| + c}{1 - |a| - 2c}\right)^2\right\}\right\}.$$

Clearly (31) holds. For $y \notin C$, we have

$$\begin{aligned} E\{V(Y_t)|Y_{t-1} = y\} &\leq 1 + E|\epsilon_t| + \left(|a| + |b|e^{-\gamma y^2}\right)|y| \\ &< 1 + E|\epsilon_t| + (|a| + c)|y| \\ &< (1-c)V(y), \end{aligned} \tag{35}$$

which shows (iii). Thus (34) admits a stationary solution with geometrically decreasing β-mixing coefficients, whenever $|a| < 1$, whatever $b \in \mathbb{R}$ and $\gamma > 0$. This is not

surprising because (34) can be interpreted as a model which can pass smoothly from an AR(1) model with parameter $a+b$ when Y_{t-1}^2 is small to an AR(1) model with parameter a when Y_{t-1}^2 is large.

From the second inequality in (35), one can see that iv) given in (33) is satisfied, and thus the stationary solution admits a finite moment of order 1, whenever $E|\epsilon_t| < \infty$. Similar arguments show that $E|Y_t|^k < \infty$ whenever $E|\epsilon_t|^k < \infty$.

5 Identification, estimation and model adequacy checking

The quasi-maximum likelihood (QML) and nonlinear least squares (NLS) estimators are widely used for the statistical inference of nonlinear time series models. For an extensive discussion of the asymptotic theory of the QML and NLS estimators in a very general framework, the reader is referred to [100] and the references therein. Of course, numerous other estimation methods are useful in nonlinear time series analysis.

We will focus on the QML estimator (QMLE) for univariate models of the form

$$Y_t = m_{\theta_0}(Y_{t-1}, Y_{t-2}, \dots) + \sigma_{\theta_0}(Y_{t-1}, Y_{t-2}, \dots)\eta_t, \tag{36}$$

where θ_0 is an unknown parameter belonging to a subset Θ of $\mathbb{R}^s$, and (η_t) is IID$(0,1)$, with η_t independent of Y_{t-i} for $i > 0$. Under these assumptions, we have

$$m_t(\theta_0) := m_{\theta_0}(Y_{t-1}, Y_{t-2}, \dots) = E(Y_t \mid Y_{t-1}, Y_{t-2}, \dots)$$

and

$$\sigma_t^2(\theta_0) := \sigma_{\theta_0}^2(Y_{t-1}, Y_{t-2}, \dots) = \mathrm{Var}(Y_t \mid Y_{t-1}, Y_{t-2}, \dots).$$

Assume that η_t has density f. Given initial values $Y_0, Y_{-1}, \dots$ the (conditional) likelihood of the observations $Y_1, \dots, Y_n$ evaluated at $\theta \in \Theta$ is equal to

$$L_n(\theta; Y_1, \dots, Y_n) = \prod_{t=1}^{n} \frac{1}{\sigma_t(\theta)} f\left(\frac{Y_t - m_t(\theta)}{\sigma_t(\theta)}\right),$$

assuming $\sigma_t(\theta) \neq 0$. This objective function is not operational because f and the initial values are generally unknown. The QML is obtained by replacing the density $f(x)$ by the $\mathcal{N}(0,1)$ density, and the conditional moments $m_t(\theta)$ and $\sigma_t^2(\theta)$ by measurable approximations $\tilde{m}_t(\theta)$ and $\tilde{\sigma}_t^2(\theta)$. One can take $\tilde{m}_t(\theta) = E_\theta(Y_t \mid Y_{t-1}, \dots, Y_1)$ and $\tilde{\sigma}_t^2(\theta) = \mathrm{Var}_\theta(Y_t \mid Y_{t-1}, \dots, Y_1)$ when these quantities are available. It is often simpler to work with approximations of the form $\tilde{m}_t(\theta) = m_\theta(Y_{t-1}, \dots, Y_1, 0, \dots)$ and $\tilde{\sigma}_t^2(\theta) = \sigma_\theta^2(Y_{t-1}, \dots, Y_1, 0, \dots)$. A QML estimator of θ_0 is defined as any measurable solution $\hat{\theta}_n$ of

$$\hat{\theta}_n = \arg\inf_{\theta\in\Theta} \tilde{Q}_n(\theta) \tag{37}$$

where, omitting several "(θ)" to lighten the notations,

$$\tilde{Q}_n(\theta) = n^{-1}\sum_{t=1}^{n}\tilde{\ell}_t, \qquad \text{and} \quad \tilde{\ell}_t = \tilde{\ell}_t(\theta) = \frac{(Y_t - \tilde{m}_t)^2}{\tilde{\sigma}_t^2} + \log\tilde{\sigma}_t^2. \tag{38}$$

The NLS estimator is obtained by assuming that the conditional variance $\tilde{\sigma}_t^2$ is constant. The existence of a solution to (37) is guaranteed when

(i) Θ is compact and the functions $\theta \to \tilde{m}_t(\theta)$ and $\theta \to \tilde{\sigma}_t^2(\theta) > 0$ are continuous.

5.1 Consistency of the QMLE

Assume that

(ii) (Y_t) is a non anticipative strictly stationary and ergodic solution of (36).

Then (m_t), (σ_t^2) and (ℓ_t), with $\ell_t = (Y_t - m_t)^2\sigma_t^{-2} + \log\sigma_t^2$, are also stationary ergodic processes. In view of the ergodic theorem (24), the theoretical criterion $Q_n(\theta) = n^{-1}\sum_{t=1}^{n}\ell_t$ thus converges almost surely to the asymptotic criterion

$$Q_\infty(\theta) = E_{\theta_0}\ell_t(\theta),$$

provided the expectation is well defined. Note that this expectation exists in $\mathbb{R}\cup\{+\infty\}$ for all θ, and in $\mathbb{R}$ for $\theta = \theta_0$, under the mild moment assumption

$$\text{(iii) } E\log^-\sigma_t^2(\theta) < \infty \text{ for all } \theta\in\Theta, \quad \text{and} \quad E\log^+\sigma_t^2(\theta_0) < \infty.$$

The initial values are often uniformly negligible, in the sense that

$$\text{(iv) } \sup_{\theta\in\Theta}\left|\ell_t - \tilde{\ell}_t\right| \to 0 \ a.s. \quad \text{as} \quad t\to\infty.$$

We then have

$$\tilde{Q}_n(\theta) - Q_n(\theta) \to 0 \quad a.s. \text{ uniformly in } \theta, \tag{39}$$

and the operational criterion $\tilde{Q}_n(\theta)$ also converges to the asymptotic criterion $Q_\infty(\theta)$.

We now need an identifiability assumption

$$\text{(v) if } \theta\neq\theta_0 \quad \text{then} \quad m_t(\theta)\neq m_t(\theta_0) \quad \text{or} \quad \sigma_t^2(\theta)\neq\sigma_t^2(\theta_0)$$

with non zero probability. Since $Q_\infty(\theta_0) = 1 + E\log\sigma_t^2(\theta_0)$ is finite, $\log x \le x-1$ for all $x>0$, and $\log x = x-1$ if and only if $x=1$, we have

$$
\begin{aligned}
Q_\infty(\theta) - Q_\infty(\theta_0) &= E_{\theta_0} \log \frac{\sigma_t^2(\theta)}{\sigma_t^2(\theta_0)} + E_{\theta_0} \left\{ \frac{Y_t - m_t(\theta)}{\sigma_t(\theta)} \right\}^2 - 1 \\
&= E_{\theta_0} \log \frac{\sigma_t^2(\theta)}{\sigma_t^2(\theta_0)} + E_{\theta_0} \left\{ \frac{m_t(\theta_0) - m_t(\theta)}{\sigma_t(\theta)} \right\}^2 \\
&\quad + E_{\theta_0} \left\{ \frac{\sigma_t(\theta_0)\eta_t}{\sigma_t(\theta)} \right\}^2 - 1 \\
&\geq E_{\theta_0} \left\{ \log \frac{\sigma_t^2(\theta)}{\sigma_t^2(\theta_0)} + \log \frac{\sigma_t^2(\theta_0)}{\sigma_t^2(\theta)} \right\} = 0
\end{aligned}
$$

with equality if and only if $m_t(\theta_0) = m_t(\theta)$ and $\sigma_t^2(\theta_0)/\sigma_t^2(\theta) = 1$ P_{θ_0}-$a.s.$ In view of (v), this latter condition is equivalent to $\theta = \theta_0$. Thus we have shown that the asymptotic criterion is minimum at the true value θ_0. This is not sufficient to claim the consistency. Indeed we have shown that

$$
\theta_0 = \arg \inf_{\theta \in \Theta} \lim_{n \to \infty} \tilde{Q}_n(\theta) \quad a.s.
$$

whereas we would like to show that

$$
\theta_0 = \lim_{n \to \infty} \arg \inf_{\theta \in \Theta} \tilde{Q}_n(\theta) \quad a.s. \tag{40}
$$

The problem can be solved as follows. Let $\theta_1 \neq \theta_0$ and $V_d(\theta_1)$ the open sphere with center θ_1 and radius $1/d$. The process $\left\{ \inf_{\theta \in V_d(\theta_1) \cap \Theta} \ell_t(\theta) \right\}_t$ is stationary and ergodic. Applying once again the ergodic theorem, we have

$$
\inf_{\theta \in V_d(\theta_1) \cap \Theta} Q_n(\theta) \geq \frac{1}{n} \sum_{i=1}^{n} \inf_{\theta \in V_d(\theta_1) \cap \Theta} \ell_t(\theta) \overset{p.s.}{\to} E \inf_{\theta \in V_d(\theta_1) \cap \Theta} \ell_t(\theta).
$$

In view of the continuity of $\ell_t(\cdot)$, the sequence $\inf_{\theta \in V_d(\theta_1) \cap \Theta} \ell_t(\theta)$ increases to $\ell_t(\theta_1)$ when $d \to \infty$. By the Beppo-Levi theorem,

$$
\lim_{d \to \infty} \uparrow E \inf_{\theta \in V_d(\theta_1) \cap \Theta} \ell_t(\theta) = E \lim_{d \to \infty} \uparrow \inf_{\theta \in V_d(\theta_1) \cap \Theta} \ell_t(\theta) = E\ell_t(\theta_1) > Q_\infty(\theta_0).
$$

Thus we have shown that for all $\theta_i \neq \theta_0$ there exists a neighborhood $V(\theta_i)$ such that

$$
\liminf_{n \to \infty} \inf_{\theta \in V(\theta_i) \cap \Theta} Q_n(\theta) > \lim_{n \to \infty} Q_n(\theta_0). \tag{41}
$$

The compact set Θ is covered by a finite number of open sets $V(\theta_1), \ldots, V(\theta_m)$ and $V(\theta_0)$, where $V(\theta_0)$ is any neighborhood of θ_0, and the neighborhoods $V(\theta_i)$ $i = 1, \ldots, m$ satisfy (41). With probability 1 we then have

$$
\inf_{\theta \in \Theta} Q_n(\theta) = \min_{i=0,1,\ldots,m} \inf_{\theta \in V(\theta_i)} Q_n(\theta) = \inf_{\theta \in V(\theta_0)} Q_n(\theta)
$$

for n large enough. Since $V(\theta_0)$ can be an arbitrarily small neighborhood of θ_0, using (39), the consistency result (40) is shown.

5.2 Asymptotic distribution of the QMLE

In addition to the previous assumptions, assume that

$$\text{(vi) } \theta_0 \text{ belongs to the interior } \overset{\circ}{\Theta} \text{ of } \Theta,$$

(vii) $\theta \to m_t(\theta)$ and $\theta \to \sigma_t(\theta)$ admit continuous third order derivatives, and

$$E \sup_{\theta \in \Theta} \left| \frac{\partial^3 \ell_t(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| < \infty \quad \forall i, j, k.$$

(viii) $I = E \frac{\partial \ell_t(\theta_0)}{\partial\theta} \frac{\partial \ell_t(\theta_0)}{\partial\theta'}$ and $J = E \frac{\partial^2 \ell_t(\theta_0)}{\partial\theta\partial\theta'}$ exist and are non singular.

Assumption (vi) is essential to obtain the asymptotic normality of the QLME. Assumption (vii) is not necessary, but can be used to easily show, by means of a Taylor expansion, that

$$\frac{\partial^2 Q_n(\theta_n^*)}{\partial\theta\partial\theta'} \to J \quad a.s. \text{ for any sequence } \theta_n^* \text{ between } \hat{\theta}_n \text{ and } \theta_0.$$

Because $\tilde{Q}_n(\theta)$ is minimized at $\hat{\theta}_n$ which, at least for n large enough, belongs to the interior of Θ, we have $\partial \tilde{Q}_n(\hat{\theta}_n)/\partial\theta = 0$. A Taylor expansion then yields

$$0 = \sqrt{n} \frac{\partial \tilde{Q}_n(\hat{\theta}_n)}{\partial\theta} = \sqrt{n} \frac{\partial \tilde{Q}_n(\theta_0)}{\partial\theta} + \left(\frac{\partial^2 \tilde{Q}_n(\theta^*)}{\partial\theta\partial\theta'} \right) \sqrt{n}(\hat{\theta}_n - \theta_0)$$

where the matrix $\left(\frac{\partial^2 \tilde{Q}_n(\theta^*)}{\partial\theta\partial\theta'} \right)$ has elements of the form $\frac{\partial^2 \tilde{Q}_n(\theta_{ij}^*)}{\partial\theta_i \partial\theta_j}$, with θ_{ij}^* between θ_0 and $\hat{\theta}_n$. The initial values being uniformly negligible, we can often show that

(ix) $\left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \ell_t(\theta_0)}{\partial\theta} - \frac{\partial \tilde{\ell}_t(\theta_0)}{\partial\theta} \right\| \xrightarrow{P} 0$ as $t \to \infty$, and that (iv) continues to hold when ℓ_t and $\tilde{\ell}_t$ are replaced by their partial derivatives up to order 3.

We deduce that for n sufficiently large $\frac{\partial^2 \tilde{Q}_n(\theta^*)}{\partial\theta\partial\theta}$ is invertible, and that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left(\frac{\partial^2 \tilde{Q}_n(\theta^*)}{\partial\theta\partial\theta'} \right)^{-1} \sqrt{n} \frac{\partial \tilde{Q}_n(\theta_0)}{\partial\theta} \overset{o_P(1)}{=} -J^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t,$$

where

$$Z_t = -2\eta_t \frac{\partial m_t(\theta_0)}{\partial \theta} + \left\{1 - \eta_t^2\right\} \frac{1}{\sigma_t^2} \frac{\partial \sigma_t^2(\theta_0)}{\partial \theta}.$$

Note that $\{Z_t, \sigma(\eta_u, u \leq t)\}_t$ is a square integrable stationary martingale difference. The central limit theorem of [19] allows to conclude that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma := J^{-1} I J^{-1}\right). \tag{42}$$

Example 11 *Let us consider once again the EXPAR(1) model (34) with* (ϵ_t) $\mathrm{IID}(0, \sigma^2)$*, adding the subscript "0" to the unknown parameters. The parameter of interest is* $\theta_0 = (a_0, b_0, \gamma_0)$*, and* σ_0^2 *can be considered as a nuisance parameter. Take the parameter space of the form* $\Theta = [-\overline{a}, \overline{a}] \times [-\overline{b}, \overline{b}] \times [\underline{\gamma}, \overline{\gamma}]$ *with* $\overline{a} \geq 1$*,* $|b_0| < \overline{b}$ *and* $0 < \underline{\gamma} < \gamma_0 < \overline{\gamma}$*, so that* $\theta_0 \in \overset{\circ}{\Theta}$*. We have* $\tilde{m}_1 = 0$ *and* $\tilde{m}_t = m_t = (a + be^{-\gamma Y_{t-1}^2}) Y_{t-1}$ *for all* $t > 1$*. Since* $\sigma_t^2(\theta) \equiv \sigma_0^2 > 0$*, the QMLE coincides with the NLS estimator. The conditions (i), (iii) and (iv) are obviously satisfied, and Example 10 shows that (ii) can be assumed when* ϵ_t *has a density* $f > 0$*. It can be shown that the identifiability condition (v) holds if and only if* $b_0 \neq 0$ *(when* $b_0 = 0$ *the parameter* γ_0 *is not identified). Thus the consistency of the QMLE is ensured when* $b_0 \neq 0$*. The conditions (vi) is obvious. The vector*

$$\sup_{\theta \in \Theta} \left| \frac{\partial \ell_t(\theta)}{\partial \theta} \right| = 2 \sup_{\theta \in \Theta} \left| \frac{(Y_t - m_t)}{\sigma^2} \begin{pmatrix} Y_{t-1} \\ Y_{t-1} e^{-\gamma Y_{t-1}^2} \\ -b Y_{t-1}^3 e^{-\gamma Y_{t-1}^2} \end{pmatrix} \right|$$

admits a finite expectation when $EY_t^2 < \infty$ *(using the fact that* $y^k e^{-\gamma y^2}$ *is bounded). Extending the argument to the third-order derivatives, we see that (vii) is satisfied and* I *and* J *exist when* $EY_t^4 < \infty$*. In view of Example 10, it suffices to assume that* $E\epsilon_t^4 < \infty$*. When* $(c_1, c_2, c_3, c_4) \neq 0$ *and* $b_0 \neq 0$*, the set* $\{y : c_1 y + c_2 y e^{-\gamma_0 y^2} + c_3 b_0 y^2 e^{-\gamma_0 y^2} = c_4\}$ *is finite. Since* Y_t *has a continuous distribution, we deduce that the components of* $\partial \ell_t(\theta_0)/\partial \theta$ *are not almost surely linearly dependent. Thus* I *and* $J = I/2$ *are invertible, and (viii) is shown. Because* $\tilde{\ell}_t = \ell_t$ *for* $t > 1$*, (ix) holds true. The asymptotic normality of the QMLE follows. Similar results have been obtained by [115], under slightly different conditions. Figure 9 summarizes the main results of a simulation experiment, in which the finite sample distribution of the NLS estimator is close to the asymptotic one.*

The constraint $b_0 \neq 0$ *is however an important restriction. Indeed we do not know the behaviour of the QMLE when the DGP is a strong AR(1), so we can not employ standard strong linearity tests, such as the Wald test.*

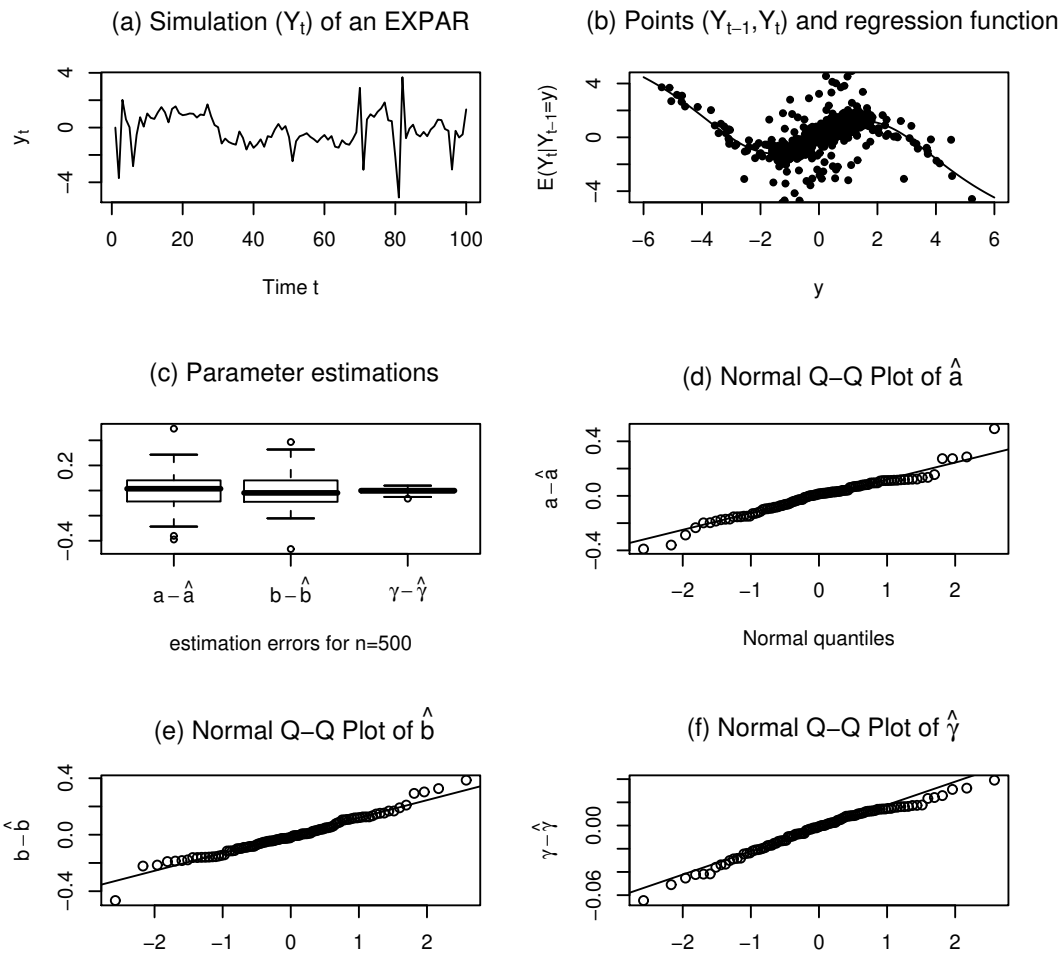


Figure 9. NLS estimates of 100 independent replications of simulations of length $n = 500$ of the EXPAR model $Y_t = \left\{-0.8 + 2\exp(-0.1Y_{t-1}^2)\right\} Y_{t-1} + \epsilon_t$, where $\epsilon_t \sim$ IID with the mixture distribution $0.9\mathcal{N}(0, 0.5^2) + 0.05\mathcal{N}(3, 1) + 0.05\mathcal{N}(-3, 1)$.

5.3 Identification and model adequacy

The information matrices I and J can be consistently estimated by their empirical counterparts

$$\hat{I} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \tilde{\ell}_t(\hat{\theta}_n)}{\partial \theta} \frac{\partial \tilde{\ell}_t(\hat{\theta}_n)}{\partial \theta'} \qquad \text{and} \qquad \hat{J} = \frac{\partial^2 \tilde{Q}_n(\hat{\theta}_n)}{\partial \theta \partial \theta'}.$$

Approximating Σ by $\hat{\Sigma} = \hat{J}^{-1}\hat{I}\hat{J}^{-1}$, the asymptotic normality (42) of the QMLE can be directly exploited to obtain asymptotic confidence region or to perform tests on the parameters. Consider for instance the null hypothesis $H_0 : R\theta_0 = r$, where R is a matrix of full row rank s_0 and r is a vector. The

Wald test rejects H_0 at the asymptotic level α when the statistic

$$\mathrm{W} = n(R\hat{\theta}_n - r)' \left\{ R\hat{\Sigma}R' \right\}^{-1} (R\hat{\theta}_n - r) > \chi^2_{s_0}(1-\alpha).$$

Such tests are generally employed to see whether the model (36) can be simplified, *i.e.* if the number s of parameters can be reduced. To see if the model (36) is sufficiently rich to take into account the dynamics of the series, practitioners often plot the residuals. Portmanteau tests based on the autocorrelations of the residuals, or of the squares of the residuals, or of any other transformation of the residuals, can be performed for model adequacy checking.

The previous steps should lead to the selection of a small set of models which possess significant estimated parameters and which pass the goodness-of-fit portmanteau tests. In general, these models are not nested and may have different parameter dimension. The choice between this models is often made by minimizing information criteria. The most popular criterion for model selection is the Akaike information criterion (AIC) proposed by [1].

5.3.1 Comparing nonlinear models with the AIC criterion

Assume that, with respect to a σ-finite measure μ, the true density of the observations $Y = (Y_1, \ldots, Y_n)$ is g, and that some candidate model gives a density $f_k(\cdot, \theta_k)$ to the observations, where θ_k is a p_k-dimensional parameter. The discrepancy between the (wrong) model and the truth can be measured by the Kullback-Leibler divergence

$$\Delta\left\{f_k(\cdot,\theta_k) \mid g\right\} = E_g \log \frac{g(Y)}{f_k(Y,\theta_k)} = E_g \log g(Y) + \frac{1}{2} d\left\{f_k(\cdot,\theta_k) \mid g\right\},$$

where

$$d\left\{f_k(\cdot,\theta_k) \mid g\right\} = -2E_g \log f_k(Y,\theta_k) = -2\int \left\{\log f_k(y,\theta_k)\right\} g(y)\mu(dy)$$

is sometimes called the Kullback-Leibler contrast. The main property of the Kullback-Leibler divergence is that $\Delta\left\{f_k(\cdot,\theta_k) \mid g\right\} \geq 0$ with equality if and only if $f_k(\cdot,\theta_k) = g$. Minimizing $\Delta\left\{f_k(\cdot,\theta_k) \mid g\right\}$ with respect to $f_k(\cdot,\theta_k)$ is equivalent to minimizing the contrast $d\left\{f_k(\cdot,\theta_k) \mid g\right\}$. Let

$$\theta_{0k} = \arg\inf_{\theta_k} d\left\{f_k(\cdot,\theta_k) \mid g\right\} = \arg\inf_{\theta_k} -2E \log f_k(Y,\theta_k)$$

be an optimal parameter for the model k (assuming that such a parameter exists). The parameter θ_{0k} being unknown, one can want to find the estimated model which minimizes

$$\mathrm{C}(k) = -2E \log f_k(Z, \hat{\theta}_{n,k}), \tag{43}$$

where the expectation is taken oven Y and Z, where Y and Z are independent and have the same distribution g, and where $\hat{\theta}_{n,k}$ is a QMLE based on Y, satisfying

$$\hat{\theta}_{n,k} = \arg\sup_{\theta_k} \log f_k(Y, \theta_k).$$

A model minimizing (43) can be interpreted as a model such that its estimated version will do globally the best job on an independent copy of the DGP.

We have

$$\mathrm{C}(k) = -2E \log f_k(Y, \hat{\theta}_{n,k}) + a_1 + a_2,$$

where

$$a_1 = -2E \log f_k(Y, \theta_{0k}) + 2E \log f_k(Y, \hat{\theta}_{n,k})$$

and

$$a_2 = -2E \log f_k(Z, \hat{\theta}_{n,k}) + 2E \log f_k(Y, \theta_{0k}).$$

The QMLE satisfies $\log f_k(Y, \hat{\theta}_{n,k}) \geq \log f_k(Y, \theta_{0k})$ almost surely. Thus a_1 can be interpreted as the average overfitting of the QMLE. Note that $E \log f_k(Y, \theta_{0k}) = E \log f_k(Z, \theta_{0k})$. Thus a_2 can be interpreted as an average cost due to the use of the estimated parameter instead of the optimal parameter, when the model is applied to an independent replication of the DGP.

It is shown in [57] that, under some regularity conditions, a_1 and a_2 are both equivalent to p_k. In this case, the AIC formula

$$\mathrm{AIC}(k) = -2 \log f_k(Y, \hat{\theta}_{n,k}) + 2p_k \tag{44}$$

is an approximately unbiased estimate of the contrast $\mathrm{C}(k)$. Model selection is then obtained by minimizing (44) over the candidate models k.

We now discuss the regularity conditions needed for a_1 and a_2 be actually equivalent to p_k. Under assumptions similar to those made in Section 5.1 and 5.2, in particular the uniqueness of the optimal parameter θ_{0k}, the estimator $\hat{\theta}_{n,k}$ converges almost surely to θ_{0k} and

$$\sqrt{n}\left(\hat{\theta}_{n,k} - \theta_{0k}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, J_k^{-1} I_k J_k^{-1}\right),$$

where

$$I_k = \lim_{n\to\infty} \frac{1}{n} \mathrm{Var} \frac{\partial}{\partial\theta} \log f_k(Y, \theta_{0k}), \quad J_k \overset{a.s.}{=} -\lim_{n\to\infty} \frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta} \log f_k(Y, \theta_{0k}).$$

Moreover, a Taylor expansion of the quasi log-likelihood yields

$$-2 \log f_k(Y, \theta_{0k}) \overset{o_P(1)}{=} -2 \log f_k(Y, \hat{\theta}_{n,k}) + \sqrt{n}\left(\hat{\theta}_{n,k} - \theta_{0k}\right)' J_k \sqrt{n}\left(\hat{\theta}_{n,k} - \theta_{0k}\right).$$

Taking the expectation of both sides, and showing that

$$En\left(\hat{\theta}_{n,k}-\theta_{0k}\right)' J_k\left(\hat{\theta}_{n,k}-\theta_{0k}\right) = \text{trace}\left\{J_k E\, n\left(\hat{\theta}_{n,k}-\theta_{0k}\right)\left(\hat{\theta}_{n,k}-\theta_{0k}\right)'\right\}$$
$$\rightarrow \text{trace}\left(I_k J_k^{-1}\right),$$

we obtain $a_1 \stackrel{o(1)}{=} \text{trace}\left(I_k J_k^{-1}\right)$. Now a Taylor expansion of the contrast yields

$$\begin{aligned} d\left\{f_k(\cdot,\hat{\theta}_{n,k}) \mid g\right\} &\stackrel{o_P(1)}{=} d\left\{f_k(\cdot,\theta_{0k}) \mid g\right\} + \left(\hat{\theta}_{n,k}-\theta_{0k}\right)' \left.\frac{\partial d\left\{f_k(\cdot,\theta) \mid g\right\}}{\partial\theta}\right|_{\theta=\theta_{0k}} \\ &\quad + \frac{1}{2}\left(\hat{\theta}_{n,k}-\theta_{0k}\right)' \left.\frac{\partial^2 d\left\{f_k(\cdot,\theta) \mid g\right\}}{\partial\theta\partial\theta'}\right|_{\theta=\theta_{0k}} \left(\hat{\theta}_{n,k}-\theta_{0k}\right) \\ &\stackrel{o_P(1)}{=} d\left\{f_k(\cdot,\theta_{0k}) \mid g\right\} + n\left(\hat{\theta}_{n,k}-\theta_{0k}\right) J_k \left(\hat{\theta}_{n,k}-\theta_{0k}\right)', \end{aligned}$$

assuming that the contrast is smooth enough, and that we can take its derivatives under the expectation sign. We deduce that

$$\begin{aligned} -2E\log f_k(Z,\hat{\theta}_{n,k}) &= E_Y d\left\{f_k(\cdot,\hat{\theta}_{n,k}) \mid g\right\} \\ &\stackrel{o_P(1)}{=} d\left\{f_k(\cdot,\theta_{0k}) \mid g\right\} + \text{trace}\left(I_k J_k^{-1}\right), \end{aligned}$$

which shows that a_2 is equivalent to a_1. Note that when $I_k = J_k$, we have $\text{trace}\left(I_k J_k^{-1}\right) = p_k$ and, in this case, the AIC(k) defined by (44) really is an approximately unbiased estimator of C(k).

6 Forecasting with nonlinear models

The complex structure of the nonlinear models is able to catch specific features of the time series, typically neglected by linear models. These specificities are referred to as *non linear features*, related for example to the nonnormality of the errors, nonlinear relationships among variables, or bimodality of the generating process. However empirical studies highlight that the good fitting results of nonlinear models do not guarantee an equally good performance in terms of forecast accuracy [39]. This often depends on the sensitivity of predictions to initial conditions and to the forecast horizon, and on some factors which are beyond the variables specified in the predictor [127]. The listed aspects have heavy impact on the generation and evaluation of forecasts from nonlinear models, and for this purpose a large amount of techniques has been proposed in the literature. However much remain to be done and further research is going on in this area. The attention here is mainly restricted to parametric nonlinear models for the conditional mean. Issues related to the volatility forecasting are also mentioned.

When dealing with forecasting one may be interested in point forecasts, interval forecasts or density forecasts. Furthermore attention must be paid to forecasts evaluation and to the opportunity of combining forecasts in order to improve the forecasting performance. Nonlinear forecasting has been addressed, among the others, in Tong [120], Granger and Teräsvirta [69], Franses and van Dijk [62], from a financial prospective, and Fan and Yao [55] which focus on nonparametric issues. For large scale comparison of forecasting performance of linear and nonlinear models, the reader is referred to Stock and Watson [108] and Marcellino [93]. For a recent survey see Timmermann[112].

6.1 Forecast generation

For most nonlinear models of the conditional mean, the generation of one-step-ahead forecasts is straightforward, but problems could arise for multi-step-ahead forecasts. The distribution of the non-linear predictors is often skewed (even when the errors in the models have a symmetric distribution) and multimodal. Differently from linear models, predictive uncertainty of nonlinear models does not necessarily grow as the *lead time* increases.

Suppose that $\Omega_t = \{Y_1, \ldots, Y_t\}$ is an observed time series and h is the lead time. The least squares predictor of Y_{t+h} is defined as

$$f_{t,h}(\Omega_t) = \arg\inf_f E[Y_{t+h} - f(\Omega_t)]^2, \tag{45}$$

where $f(\cdot)$ denotes a measurable function over Ω_t. It is easy to show that

$$f_{t,h}(\Omega_t) = E[Y_{t+h} \mid \Omega_t] \equiv Y_t(h).$$

When $f(\cdot)$ is a linear function, $Y_t(h)$ has some optimal properties in terms of predictive accuracy and variability as shown in [24].

The same results do not always hold when $f(\cdot)$ has a nonlinear structure [55].

In this framework Tong [118] originally suggests to generate forecasts through the naïve method which allows to catch the skeleton of the data generating process setting to zero the error term.

Example 12 *Consider the SETAR$(k; p_1, p_2, \ldots, p_k)$ model represented as*

$$Y_t = \sum_{j=1}^{k} (\nu^{(j)} + \sum_{i=1}^{p_j} a_i^{(j)} Y_{t-i} + \epsilon_t^{(j)}) \mathbf{1}_{\{r_{j-1} < Y_{t-d} \leq r_j\}},$$

where $\epsilon_t^{(j)} = \sigma^{(j)}\eta_t$ *and* (η_t) *is a noise with unit variance. Let* $h \leq d$*. The* h*-step-ahead naïve prediction of the SETAR model,* $\hat{Y}_t^n(h)$*, is given as a combination of the estimated and observed values*

$$\hat{Y}_t^n(h) = \sum_{j=1}^{k} \left\{ \nu^{(j)} + \sum_{i=1}^{h-1} a_i^{(j)} \hat{Y}_t^n(h-i) + \sum_{i=h}^{p_j} a_i^{(j)} Y_{t+h-i} \right\} \mathbf{1}_{\{r_{j-1} < Y_{t+h-d} \leq r_j\}}.$$

However this naïve approach, when used for multi-step-ahead forecasts, generates biased forecasts and can be misleading [31] [86]. In order to evaluate the predictive accuracy, *it is always useful to have available the predictive distribution* [120] which corresponds, in this context, to the conditional distribution of Y_{t+h} given Ω_t. Brown and Mariano [31] underline the role of the forecasts generated by taking expectation with respect to the know conditional distribution. They refer to this as the closed-form-forecast.

The knowledge of the predictive distribution allows to obtain the conditional expectation of Y_{t+h}, $E[Y_{t+h}|\Omega_t]$, using the closed form forecast

$$Y_t(h) = \int_{-\infty}^{\infty} Y_{t+h} \; g(Y_{t+h} \mid \Omega_t) \; dY_{t+h},$$

where $g(Y_{t+h} \mid \Omega_t)$ is the distribution of Y_{t+h} conditioned upon the past information Ω_t.

In general, deriving the analytic expression of nonlinear forecast, when $h > 1$, is a challenging task. However, by using the so-called Chapman-Kolmogorov relationship, exact least squares multi-step-ahead forecasts for general nonlinear AR models can, in principle, be obtained through complex numerical integration.

Early examples of this approach are Tong and Moeanaddin [122] that use the recursive formula of the Chapman-Kolmogorov relation to obtain h-step forecasts from threshold models and Al Qassen and Lane [2] that use it for the EXPAR models. Later on, de Gooijer and de Bruin [66], derive the conditional probability density function (p.d.f.) for h-step-ahead forecasts of first order SETAR models with Gaussian errors. The procedure can be generalized to higher-order models.

Moeanaddin [96], assuming the Gaussianity of the errors, proposes to generate the multi-step ahead forecasts of a SETAR(2;1,1) as a weighted mean of the naïve predictions obtained from each regime. Let $\hat{Y}_t^1(h)$ and $\hat{Y}_t^2(h)$ be the h-step-ahead forecasts generated from the first and the second regime. We

have

$$\hat{Y}_t^1(h) = \nu^{(1)} + a_1^{(1)}\hat{Y}_t(h-1) \quad \text{and} \quad \hat{Y}_t^2(h) = \nu^{(2)} + a_1^{(2)}\hat{Y}_t(h-1),$$

where $\hat{Y}_t(h)$ is obtained as the weighted mean

$$\hat{Y}_t(h) = p_{h-1}\hat{Y}_t^1(h) + (1-p_{h-1})\hat{Y}_t^2(h)$$

and p_{h-1} is selected through the cumulative distribution function of the standardized normal

$$p_{h-1} = \Phi\left\{\frac{r_1 - \hat{Y}_t(h-1)}{\hat{\sigma}_t(h-1)}\right\}.$$

In this regard Potter [101] observes that when the forecast horizon exceeds the length of the delay lag d, the generation of forecasts from threshold models may require the use of simulations (which could give a reasonable approximation of the empirical distribution of Y_{t+h} too). In this case the results given in Moeanaddin [96] and in de Gooijer and de Bruin [66] can only approximate the p.d.f. of predictors.

This issue was investigated in Amendola and Niglio [3] who derive the conditional p.d.f. of Y_{t+h} and the exact form of the forecast $\hat{Y}_t^c(h) = E[Y_{t+h} \mid \Omega_t]$, when Y_t follows a SETAR model under the Gaussian assumption on the error term, and assuming known parameters. The local autoregressive structure makes the Gaussian hypothesis reasonable in the SETAR context. The generation of forecasts can however be affected by several other aspects, mainly related to the *threshold variable*, and the *threshold delay*, which have important implications for the form of the predictor and for its distribution [4].

Example 13 *For a SETAR(2;1,1) model with $d = 1$, define the closed form predictor $\hat{Y}_t^c(h)$ at $h = 2$ by*

$$\begin{aligned}\hat{Y}_t^c(2) = &\int_{-\infty}^{r}\left(\nu^{(1)} + a_1^{(1)}Y_{t+1}\right) g(Y_{t+1} \mid \Omega_t)\, dY_{t+1} + \\ &+ \int_{r}^{\infty}\left(\nu^{(2)} + a_1^{(2)}Y_{t+1}\right) g(Y_{t+1} \mid \Omega_t)\, dY_{t+1} \\ &= \left\{\nu^{(1)} + a_1^{(1)}Y_t(1)\right\}\lambda + \left\{\nu^{(2)} + a_1^{(2)}Y_t(1)\right\}(1-\lambda) + (a_1^{(2)} - a_1^{(1)})\gamma,\end{aligned}$$

where

$$\lambda = \int_{-\infty}^{\{r_1 - Y_t(1)\}/\sigma_t(1)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$
$$\gamma = \frac{\sigma_t(1)}{\sqrt{2\pi}} \exp\left(-\frac{\{r_1 - Y_t(1)\}^2}{2\sigma_t^2(1)}\right)$$
$$\sigma_t(1) = \sigma^{(1)} \mathbf{1}_{\{Y_t \leq r_1\}} + \sigma^{(2)} \mathbf{1}_{\{Y_t > r_1\}}$$

and

$$Y_t(1) = \left(\nu^{(1)} + a_1^{(1)} Y_t\right) \mathbf{1}_{\{Y_t \leq r_1\}} + \left(\nu^{(2)} + a_1^{(2)} Y_t\right) \mathbf{1}_{\{Y_t > r_1\}}.$$

It can be shown that $\hat{Y}_t^c(2) = Y_t(2) = E(Y_{t+2} \mid \Omega_t)$ *under the Gaussian assumption.*

The derivation of analytical distributions of non-linear predictors is closely related to the structure of each class of model, and the difficulty increases for large values of the forecasting horizon h.

A widely used approach to generate multi-step-ahead forecasts is based on numerical techniques [69]. A simulation based forecast method is the Monte Carlo method which, in the framework of Example 12, approximates the conditional expectation $Y_t(h)$ for $h = 2, 3, \ldots$ by

$$\hat{Y}_t(h) = \frac{1}{M} \sum_{\ell=1}^{M} \left[\sum_{j=1}^{k} \left\{ \nu^{(j)} + \sum_{i=1}^{p_j} a_i^{(j)} \hat{Y}_t(h-i) + \sigma^{(i)} \eta_\ell \right\} \mathbf{1}_{\{r_{j-1} < \hat{Y}_t(h-d) \leq r_j\}} \right]$$

where the 1-step-ahead forecast is explicitly given by

$$\hat{Y}_t(1) = Y_t(1) = \sum_{j=1}^{k} \left\{ \nu^{(j)} + \sum_{i=1}^{p_j} a_i^{(j)} Y_{t+1-i} \right\} \mathbf{1}_{\{r_{j-1} < Y_{t+1-d} \leq r_j\}},$$

the convention $\hat{Y}_t(k) = Y_{t+k}$ is used for all $k \leq 0$, the number of Monte Carlo replications is M, and $\eta_1, \ldots, \eta_M$ is a sample from the presumed distribution of the errors. In practice this distribution is almost always assumed to be normal.

A closely connected alternative is the Bootstrap approach. In this case the added error terms are selected by resampling from the estimated residuals and no assumption on the errors distribution is needed. This leads this latter approach to be preferred. The h-step-ahead Bootstrap forecast is given by

$$\hat{Y}_t(h) = \frac{1}{B} \sum_{\ell=1}^{B} \left[\sum_{j=1}^{k} \left\{ \nu^{(j)} + \sum_{i=1}^{p_j} a_i^{(j)} \hat{Y}_t(h-i) + \sigma^{(i)} \eta_\ell^* \right\} \mathbf{1}_{\{r_{j-1} < \hat{Y}_t(h-d) \leq r_j\}} \right]$$

where B is the number of Bootstrap replicates, and $\eta_1^*, \ldots, \eta_B^*$ is a sample from the empirical distribution of the residuals $\hat{\eta}_1, \ldots, \hat{\eta}_n$.

The simulation-based approaches lead to easily generate non-linear multi-step-ahead forecasts and the popularity of these methods was growing up due to the increasing computing power of Pcs. However they only give an approximation of the conditional expectation.

6.2 Interval and density forecasts

In order to investigate on the accuracy of point forecasts or to evaluate the performance of a single candidate model it could be useful to compute interval forecasts, which are generally calculated as a symmetric interval around the mean [32]. In the nonlinear domain, where the predictors are often characterized by asymmetric and multimodal distributions, this procedure can be unsatisfactory. Finding the entire forecast density can be useful when a single interval may no longer provide an adequate summary of the expected future. In this context interesting proposals include fan charts [124] and forecast regions [80].

The forecast regions can be differently defined in relation to the shape of the forecast density function. In particular in order to construct a $100(1-\alpha)\%$ forecast region R_α, Hyndman has suggested three different approaches:

1) in presence of a symmetric and unimodal distribution

$$R_\alpha = [\mu_{t|h} - w_{\alpha,h},\ \ \mu_{t|h} + w_{\alpha,h}]$$

where $\mu_{t|h}$ is the mean of the distribution $g_{t,h}(y)$ of Y_{t+h} given Ω_t;

2) with asymmetric and unimodal distribution

$$R_\alpha = [Q_{t|h}(\alpha/2),\ \ Q_{t|h}(1-\alpha/2)]$$

where $Q_{t|h}(\alpha/2)$ is the $\alpha/2-$quantile of $g_{t,h}(y)$;

3) in the presence of an asymmetric and multimodal distribution

$$R_\alpha = \{y : g_{t,h}(y) \geq g(\alpha)\},$$

with $g(\alpha)$ such that $\Pr(Y_{t+h} \in R_\alpha \mid \Omega_t) = 1-\alpha$. The latter forecast region is called an High Density Region (HDR) [81].

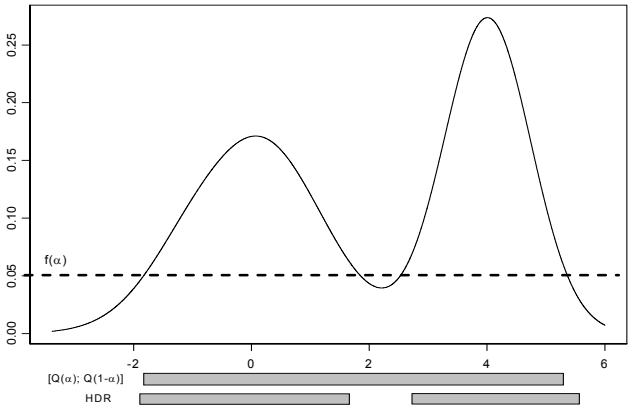


Figure 10. Two different $100(1-\alpha)\%$ forecast regions

The use of these graphical summaries has grown rapidly in recent years as density forecasts have become relatively widely used. However the need of evaluating the best predictive density requires new instruments and a few proposal have been done in the last years mainly addressing interval and density forecast evaluation. The issue is largely concerned with testing the null of correct dynamic specification. Diebold, Gunther and Tay [49] (hereafter DGT) first suggest to use the Probability Integral Transform (PIT) to evaluate the predictive density.

Let $g_t(y|\Omega_t)$ be the sequence of conditional densities of a time series Y_t and let $p_t(y|\Omega_t)$ be the corresponding sequence of 1-step-ahead density forecasts. The PIT is defined as:

$$z_t = \int_{-\infty}^{y_t} p_t(u)du = P_t(y) \tag{46}$$

If the forecasting model is correctly specified, then $p_t(y) = g_t(y)$ and in this case the sequence z_t is i.i.d. $U(0,1)$. DGT illustrate their forecasts evaluation mainly through graphical tools used even for investigate on uniformity and independence. Within this framework different tests have been proposed ([78]; [10]; [17]) that have been recently reviewed in Corradi and Swanson [42]. Clements et al. [38] pointed out, in a simulation exercise, how these tests may have negligible power to indicate the misspecification of the linear forecast density. The above procedures can be also extended for evaluate multi-step ahead density forecasts.

If we are interesting in generating the optimal point forecasts, the volatility prediction does not play any role in the forecast generation, unless the conditional mean depends directly on the conditional variance [11].

Example 14 *The GARCH-M [53] allows to include the time varying conditional variance in the equation for the mean*

$$Y_t = \nu + \sum_{i=1}^{p} a_i Y_{t-i} + \theta g(\sigma_t^2) + \epsilon_t,$$

where $g(\sigma_t^2)$ is a function of the conditional variance of the error term $\epsilon_t = \sigma_t \eta_t$, which can follow a GARCH specification. The generation of h-step-ahead forecasts of Y_{t+h} will require the prediction of future values of the conditional variance.

However, the volatility forecasts are absolutely relevant for assessing the uncertainty of the levels predictions. While the unconditional forecast error might not be affected by a time varying conditional variance, the conditional squared forecast error, $E[\epsilon_{t+h|t}^2|\Omega_t]$, is varying over time. The convergence of the mean squared forecast error (MSFE) to the unconditional variance is no more monotone [62]. Furthermore, in order to compute evaluating forecasts criteria, such as the forecast intervals, h-step-ahead forecasts for the conditional variance are required. The computation of analytic forecasts for the conditional variance can be straightforward for many GARCH-type models. The 1-step-ahead volatility forecasts depend explicitly upon the information up to time t, Ω_t. Multi-step-forecasts can then be obtained recursively (given knowledge of the model specification and of the parameters values). A recent survey on the field is [7].

Example 15 *Considering the GARCH(1,1) model*

$$\epsilon_t = \sigma_t \eta_t, \qquad \sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

the optimal h-step-ahead forecasts of the conditional variance can be computed recursively from

$$\hat{\sigma}_t^2(h) = \omega + \alpha_1 \hat{\epsilon}_{t+h-1}^2 + \beta_1 \hat{\sigma}_t^2(h-1),$$

for $h = 1, 2, \ldots,$ with $\hat{\epsilon}_{t+h-1}^2 = \hat{\sigma}_t^2(h-1)$ when $h > 1$, and the initial values

$$\hat{\epsilon}_t^2 = \epsilon_t^2, \qquad \hat{\sigma}_t^2(0) = \sum_{i=0}^{t-2} \beta^i (\omega + \alpha \epsilon_{t-i-1}^2).$$

Whereas forecasting conditional volatility is fairly straightforward, evaluating the forecasting performance is a more challenging task. Important contributions in assessing the correctness of out-of-sample conditional and unconditional interval predictions are Christoffersen [35] and Christoffersen and Diebold [36]. The main problem is that the volatility cannot be directly observed and hence loss functions such as the MSFE cannot be used unless a suitable proxy of the conditional variance is defined. A recent work on the appropriateness of certain loss functions in volatility forecast evaluation is due to Patton [99]. Common approaches are to use the squared returns, and as more recently suggested, to refer to the realized volatility concept [6]. The attention to volatility forecasts is mainly due to their important role in some application areas such as the financial market. Asset pricing and risk management are based on measuring and forecasting volatility. In this setting, the increasing attention paid to volatility forecasting is also due to the impact that accurate measures of volatility are required for computing measures of financial risk such as the Value at Risk (VaR).

6.4 Forecast combination

In order to improve the accuracy of the forecasts, combinations of different predictors have been used in forecasting from linear and nonlinear models. The literature on forecasts combination dates back to the seminal paper by Bates and Granger [14]. Given a forecast horizon h, the aim of forecasts combination is to find an optimal vector of weights such that the new predictor performs better than the single candidate models, according to an appropriately chosen loss function.

Formally, let Y_t, $t = 1, \ldots, n$, be an observed time series generated by a stationary stochastic process $\{Y_t\}$, and let $\hat{Y}_t^i(h)$ be the forecast of Y_{t+h} generated by the i-th model among k different models ($i = 1, \ldots, k$). The basic idea is to find a linear combination

$$\tilde{Y}_t(h) = \sum_{i=1}^{k} w_i \hat{Y}_t^i(h)$$

($0 \leq w_i \leq 1$, $\sum_{i=1}^{k} w_i = 1$) which performs better than the single candidate models according to an appropriately chosen criterion function e.g. the MSFE

$$E(\tilde{Y}_t(h) - Y_{t+h})^2.$$

In the original approach by Bates and Granger [14] a convexity constraint was imposed on the combination while Granger and Ramanathan [68] showed

how the accuracy of the combined predictor can be improved removing the convexity constraint and adding a constant term to the combination. In this way, even if the candidate predictors are biased their combination can still yield an unbiased predictor. A recent field of research refers to the combination of density forecast. The aim is to link two relevant aspects of the forecasting: the estimation of the density forecast and the forecasts combination. Until now few works have been published on the topic ([125], [71], [12]) which also appears to be promising to deal with nonlinear features. For instance, it is well known that when the single densities are Gaussian, the mixture can have an asymmetric behavior and/or heavy tails, allowing data with a wide range of skewness and kurtosis. Further, under proper conditions on the parameters of the densities, the mixture can even be multimodal, which is often the case in presence of nonlinearities.

7 Algorithmic aspects

Nonlinear time series analysis often makes use of iterative simulation techniques and optimization procedures. We begin with computational techniques based on simulations of Markov chains. We review MCMC methods for the bayesian inference of nonlinear time series models. In particular, an hybrid algorithm, combining Metropolis-Hastings and Gibbs steps, is derived for fitting a STAR model. We also present several algorithms used to fit models driven by hidden Markov chains. An application to the French CAC 40 stock market index is proposed.

7.1 MCMC methods

The Markov Chain Monte Carlo (MCMC) methods, in particular the Metropolis-Hastings algorithm and the Gibbs sampling, enable to simulate an ergodic Markov chain whose invariant distribution is a specified distribution. MCMC methods have become the numerical techniques of choice for many Bayesians because these methods are extremely powerful to simulate complicated posterior distribution. Most of the applied data analysts also employ MCMC algorithms for their ability the fit highly complex probability models.

7.1.1 The Metropolis-Hastings algorithm

Let $P(\theta)$ be a distribution with support $E \subset \mathbb{R}^d$. The "target" distribution P only needs to be specified up to a constant of proportionality. This is particularly interesting in a Bayesian framework when the target distribution is the

posterior distribution

$$\pi(\theta|X) = \frac{f(X \mid \theta)\pi(\theta)}{f(X)},$$

where $\pi(\theta)$ denotes the prior distribution of the parameter θ and $f(X \mid \theta)$ is the likelihood of the data X. Interestingly, the Metropolis algorithm, introduced by [94] and extended by [76], does not require the specification of the complicated constant $f(X) = \int f(X \mid \theta)\pi(\theta)d\theta$.

Given an arbitrary Markov transition kernel $Q(\cdot, \cdot)$, the Metropolis algorithm generates a Markov chain $\{\theta^{(t)}\}_{t\in\mathbb{N}}$, with state space E and invariant distribution $P(\theta)$ as follows:

(1) Choose an initial value $\theta^{(0)}$ in E.
(2) For $t = 0, 1, \ldots$
  (a) Generate a candidate value θ^* for the next state $\theta^{(t+1)}$, from the "proposal" distribution $Q(\theta^{(t)}, \cdot)$.
  (b) Compute the ratio

$$r = \frac{P(\theta^*)}{P(\theta^{(t)})} \frac{Q(\theta^*, \theta^{(t)})}{Q(\theta^{(t)}, \theta^*)}.$$

  (c) With probability $\min(r, 1)$, accept θ^* as the new state so that $\theta^{(t+1)} = \theta^*$, otherwise reject the candidate so that the chain remains at $\theta^{(t+1)} = \theta^{(t)}$.

The kernel must satisfy certain regularity conditions in order to guarantee ergodicity of the chain (see Section 4.4 for the concept of ergodicity). Otherwise the choice of $Q(\cdot, \cdot)$ is arbitrary, which is not very surprising if we realize that many Markov processes may have the same invariant probability P. Because the algorithm depends on P only through the ratio $P(\theta^*)/P(\theta^{(t)})$, as was already mentioned, it suffices that P be defined as an unnormalized probability. It is also worth noting that, although the bivariate distribution of $(\theta^{(t)}, \theta^{(t+1)})$ is not continuous ($\theta^{(t)}$ being equal to $\theta^{(t+1)}$ with nonzero probability), it is possible that the marginal distribution P be continuous.

Example 16 *The following example is artificial, and useless for the applications, but allows a better understanding of the algorithm. Assume a discrete target distribution such that* $P(0) = 1 - P(1) = p$. *Take a transition kernel such that* $Q(0, 1) = q \neq 0$ *and assume, for simplicity, that* $Q(1, 0) = Q(0, 1)$. *We have*

$$\begin{aligned} p(0,1) &:= P(\theta^{(t+1)} = 1 \mid \theta^{(t)} = 0) = P(\theta^{(t+1)} = 1,\, \theta^* = 1 \mid \theta^{(t)} = 0) \\ &= P(\theta^{(t+1)} = 1 \mid \theta^{(t)} = 0,\, \theta^* = 1) P(\theta^* = 1 \mid \theta^{(t)} = 0) \\ &= \min\{(1-p)/p, 1\}\, q. \end{aligned}$$

Similarly $p(1, 0) := P(\theta^{(t+1)} = 0 \mid \theta^{(t)} = 1) = \min\{p/(1 - p), 1\}\, q$. *The invariant distribution satisfying* $P(0) = \{1 - p(0, 1)\}P(0) + p(1, 0)P(1)$ *with* $P(1) = 1 - P(0)$,

it is then easy to check that $P(0) = 1 - P(1) = p$, whatever the value of q. Of course, when q is very small, the chain $\theta^{(t)}$ stays at the same state for a long time, and the empirical frequency of the state i ($i = 0$ or 1) is likely to converge to $P(i)$ very slowly.

Now consider a slightly more elaborate example.

Example 17 *Consider the smooth transition autoregressive model*

$$Y_t = a_0 + a_1 Y_{t-1} + \frac{(b_0 - a_0) + (b_1 - a_1)Y_{t-1}}{1 + \exp\{-\gamma(Y_{t-1} - c)\}} + \epsilon_t, \quad \epsilon_t \sim \mathrm{IID}(0, \sigma^2). \tag{47}$$

We assume $\gamma > 0$, so that the regression function $E(Y_t \mid Y_{t-1} = y)$ smoothly changes from $a_0 + a_1 y$ to $b_0 + b_1 y$ when y varies from $-\infty$ to $+\infty$. Figure 11 displays a simulated trajectory of the model.

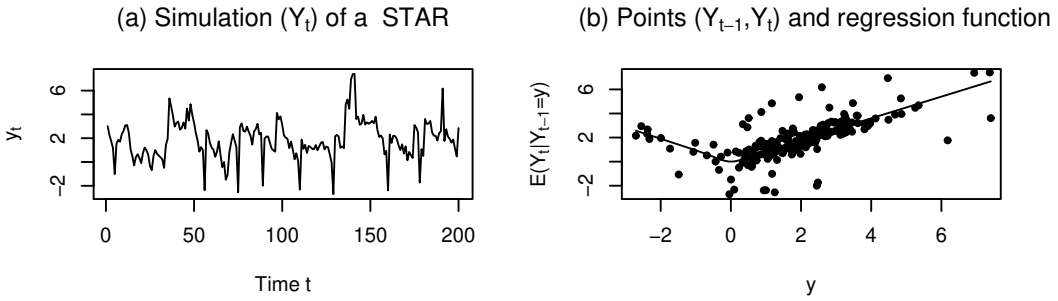


Figure 11. A simulation of length $n = 200$ of the STAR model (47) with $a_0 = b_0 = c = 0$, $a_1 = -0.95$, $b_1 = 0.9$, $\gamma = 5$, and (ϵ_t) as in Figure 9.

The **tsDyn** *package of the statistical software R (see http://cran.r-project.org/) contains a function called* `lstar()`*, which uses a standard frequentist approach to fit STAR models:*

```
> # fit a STAR model to the time series y
> fittedstar<-lstar(y,m=1,thDelay=0,control=list(maxit=3000))
> names(fittedstar$coefficients)<-c("a0","a1","b0","b1","gamma","c")
> summary(fittedstar)
Fit:
residuals variance = 1.183,  AIC = 349, MAPE = 116.4%

Coefficient(s):
       Estimate  Std. Error  t value Pr(>|t|)
a0    -0.115184    0.097962  -1.1758  0.23967
a1    -1.029256    0.068614 -15.0007  < 2e-16 ***
b0    -0.068118    0.049227  -1.3838  0.16643
b1     0.922659    0.016743  55.1081  < 2e-16 ***
gamma  9.892915    5.974123   1.6560  0.09773 .
c      0.042477    0.175698   0.2418  0.80896
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Non-linearity test of full-order LSTAR model against full-order AR model
 F = 403.39 ; p-value = 7.2136e-82
```

The function `metrop()` *of the R package* **mcmc** *(and also the function* `MCMCmetrop1R()` *of the package* **MCMCpack***) can be employed to perform a Bayesian analysis.*

Assume that under the prior distribution, all the parameters are independent, c is fixed to 0, a_0, a_1, b_0, *and* b_1 *are* $\mathcal{N}(0,1)$ *distributed,* γ *follows an exponential distribution with rate 1, and* σ^2 *follows the inverse gamma distribution with shape and scale parameters equal to 1. We begin to specify (up to an additive constant) the unnormalized log-density of the posterior distribution:*

```
> logpost<-function(theta, y){
> n<-length(y);a0<-theta[1];a1<-theta[2];b0<-theta[3];b1<-theta[4]
> gamma<-theta[5]; c<-0; sig2<-theta[6]
> if ( gamma<=0 ) return(-Inf); if ( sig2<=0 ) return(-Inf)
> logi <- 1/(1+exp(-gamma*(y[1:(n-1)]-c)))
> ychap <- (a0+a1*y[1:(n-1)])*(1-logi) + (b0+b1*y[1:(n-1)])*logi
> loglike <- -sum((y[2:n]-ychap)^2/(2*sig2))-(n-1)*log(sig2)/2
> logprior<- -(a0^2+a1^2+b0^2+b1^2)/2 - gamma -2*log(sig2)-1/sig2
> return(loglike+logprior)}
```

The following commands allow to generate a Metropolis algorithm with the random walk kernel transition $Q(x,y) \sim \mathcal{N}(x, \texttt{scale}\, I_6)$*, and with the output of the* `lstar()` *as starting value* $\theta^{(0)}$*.*

```
> theta.init0 <- {as.vector(c(fittedstar$coefficients[1:5],
+                   var(residuals(fittedstar),na.rm=T)))}
> mcmcresults <- metrop(logpost, theta.init0, 3000, scale=0.02, y=y)
> mcmcresults$accept
> chaine<-ts(mcmcresults$batch)
> dimnames(chaine)[[2]] <- c("a0","a1","b0","b1","gamma","sigma2")
> plot(chaine, main="Markov chain silumated by Metropolis")
```

The parameter `scale` *is very important for the performance of the Metropolis algorithm. High values of* `scale` *result in low acceptance rates (the candidate is likely to fall into low-density areas of the posterior distribution, and thus to be rejected, with high probability). Small values of* `scale` *result in high acceptance rate and slow movements of the chain (in this case the chain is said to be "poorly mixing"). Here the acceptance rate is* `mcmcresults$accept=0.37`*, which can be found acceptable by practitioners. Figure 12 shows however that the Markov chain generated by the algorithm does not seem to have reached its equilibrium, the component* `gamma` *moving too slowly.*

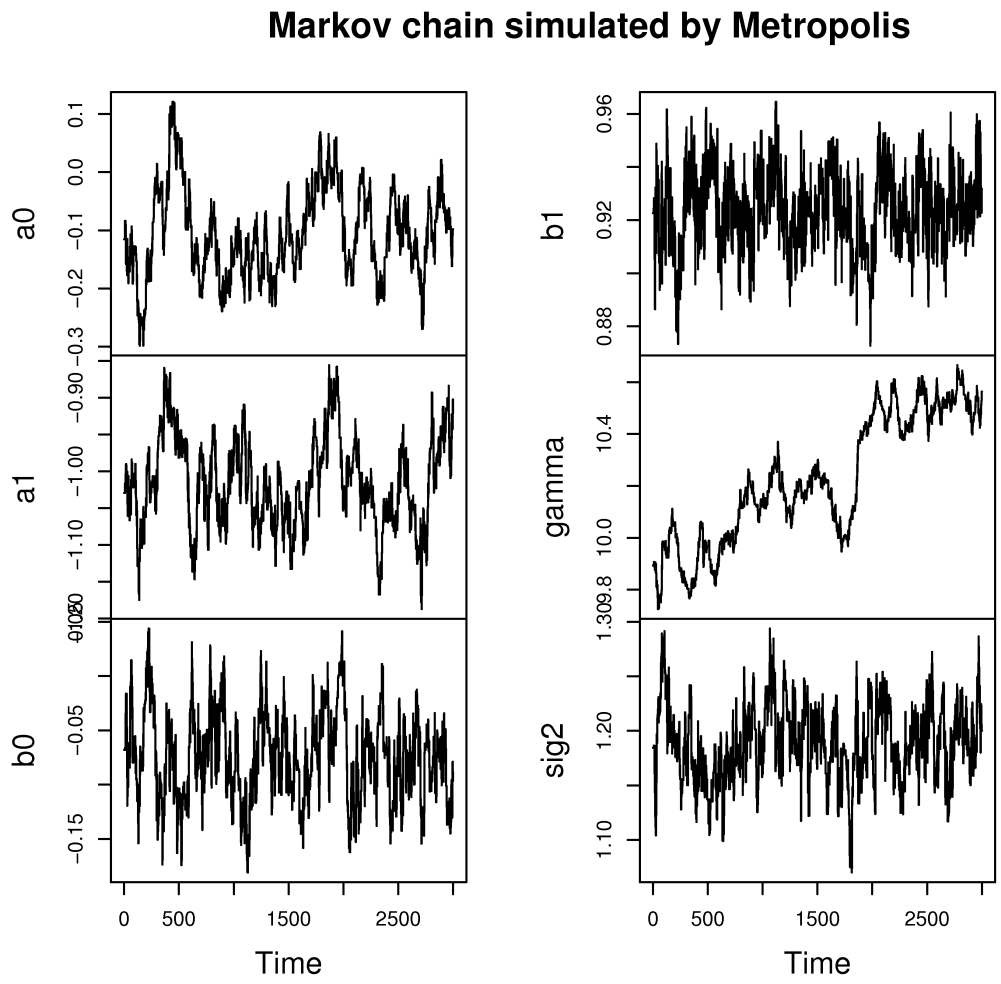


Figure 12. The Markov chain generated by the algorithm does not look stationary.

Considering the 3 000 *previous simulations as "burn-in" values, generating a longer trajectory of* 30 000 *values, and tuning the scale parameter, the following command*

```
> mcmcresults <- {metrop(mcmcresults, nbatch=30000,
+                  scale=c(0.03,0.03,0.03,0.03,0.4,0.02),y=y)}
```

generates a more satisfactory Markov chain (see Figure 13). Of course, the time series plots of Figure 13 are not sufficient to successfully diagnose "convergence" (see e.g. *[30]). Additional diagnostic tools are available in the R package* **MCMCpack**. *The empirical marginal distribution of the Markov chain generated by the Metropolis algorithm can then be used to approximate the posterior distribution of parameters (see Figure 14).*

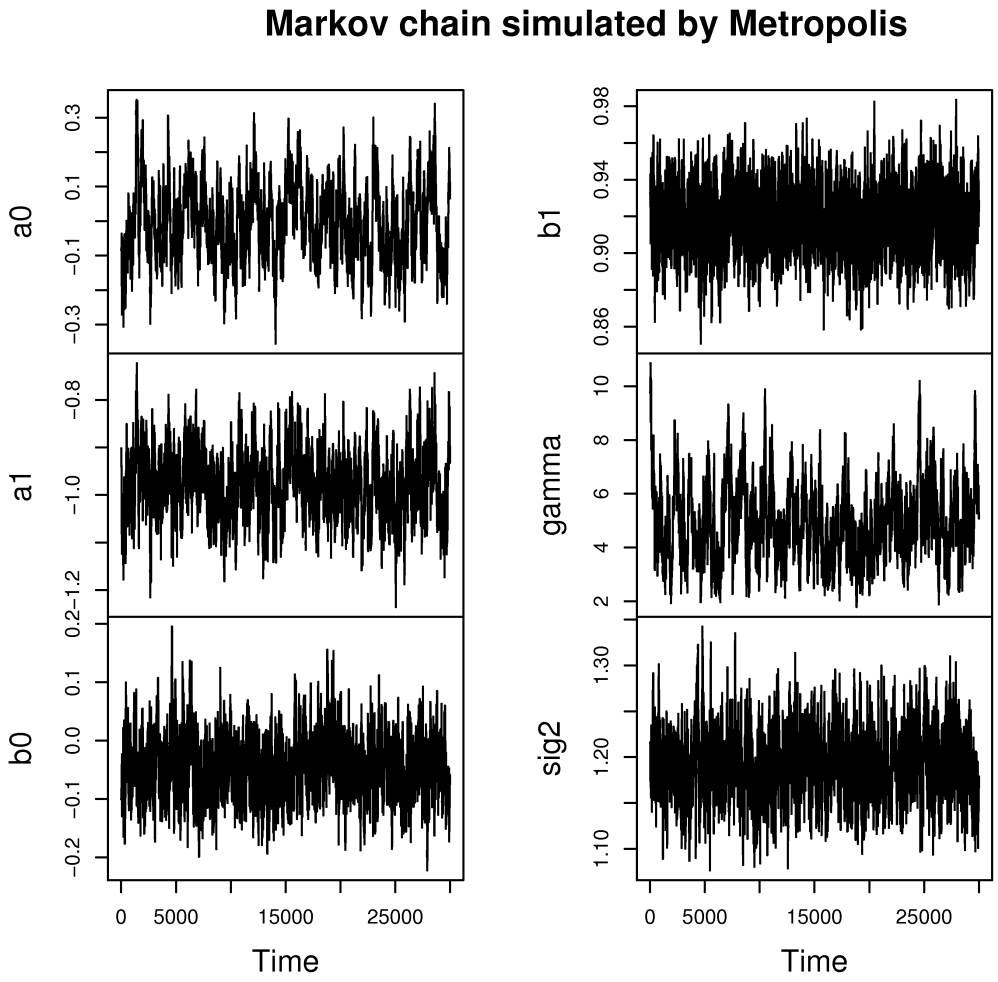


Figure 13. The trajectories do not display obvious nonstationarities.

7.1.2 The Gibbs algorithm

The Gibbs sampler, proposed by [63], allows to generate a Markov chain $\left\{\theta^{(t)}\right\} = \left\{\theta_1^{(t)}, \ldots, \theta_d^{(t)}\right\}'$ on a product state space $E = E_1 \times \cdots \times E_d$, with target distribution $P(\cdot)$. It is assume that $P(\cdot)$ has the support E, and that it is the law of a random vector $\theta = (\theta_1, \ldots, \theta_d)'$, where each θ_i has a distribution P_i with support E_i. For $i = 1, \ldots, d$, let $P_i(\cdot \mid \theta_{-i})$ be the "full conditional distribution" of θ_i given $\{\theta_j,\, j \neq i\}$. The Gibbs sampler proceeds as follows:

(1) Choose an initial value $\theta^{(0)}$ in E.
(2) For $t = 0, 1, \ldots$, generate $\theta^{(t+1)} = \left\{\theta_1^{(t+1)}, \ldots, \theta_d^{(t+1)}\right\}'$ by drawing from the lower-dimensional distributions
   (a) $\theta_1^{(t+1)} \sim P_1\left(\cdot \mid \theta_2^{(t)}, \ldots, \theta_d^{(t)}\right)$,
   (b) $\theta_i^{(t+1)} \sim P_i\left(\cdot \mid \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_d^{(t)}\right)$ for $i = 2, \ldots, d-1$,
   (c) $\theta_d^{(t+1)} \sim P_d\left(\cdot \mid \theta_1^{(t+1)}, \ldots, \theta_{d-1}^{(t+1)}\right)$.

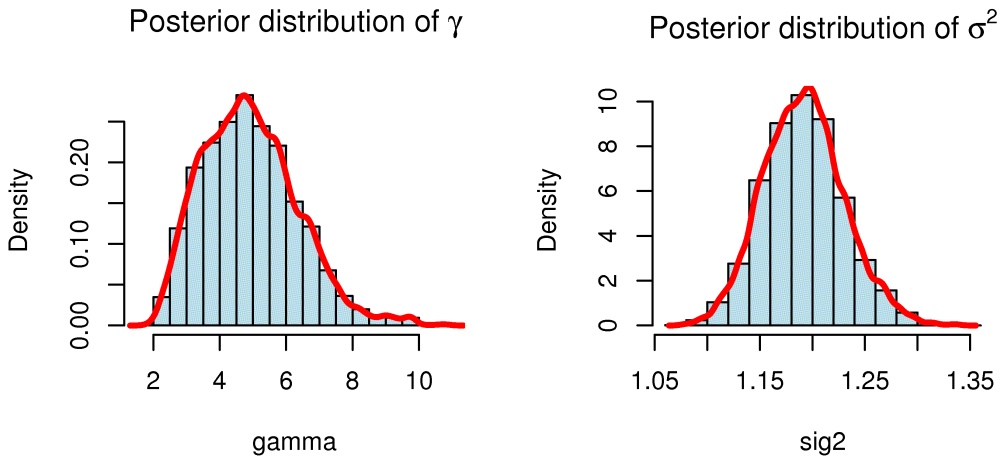


Figure 14. Histogram and kernel estimates of the posterior distributions of the parameters of the STAR model (47).

Example 18 *Let us continue with the STAR(1) model of Example 17. When γ is given, the model can be written as a linear model of the form $Y_t = Z'_{t-1}\beta + \epsilon_t$, where ϵ_t is iid $\mathcal{N}(0, \sigma^2)$, and*

$$Z_t = \begin{pmatrix} 1 - G_t(\gamma) \\ Y_t \{1 - G_t(\gamma)\} \\ G_t(\gamma) \\ Y_t G_t(\gamma) \end{pmatrix}, \quad \beta = \begin{pmatrix} a_0 \\ a_1 \\ b_0 \\ b_1 \end{pmatrix}, \quad G_t(\gamma) = \frac{1}{1 + \exp(-\gamma Y_t)}.$$

With the notation

$$\Sigma_n = \sigma^2 \left(\sum_{t=2}^{n} Z_{t-1} Z'_{t-1} \right)^{-1}, \qquad \hat{\beta} = \left(\sum_{t=2}^{n} Z_{t-1} Z'_{t-1} \right)^{-1} \left(\sum_{t=2}^{n} Y_t Z_{t-1} \right)$$

and other obvious notations, the following full conditional distributions are then explicitly given by (see e.g. *[104]):*

$$P_1(\beta \mid Y, \gamma, \sigma^2) \sim \mathcal{N}\left\{ \left(I_4 + \Sigma_n^{-1}\right)^{-1} \Sigma_n^{-1} \hat{\beta}, \ \left(I_4 + \Sigma_n^{-1}\right)^{-1} \right\},$$
$$P_2(\sigma^2 \mid Y, \beta, \sigma^2) \sim \mathcal{IG}\left(1 + (n-1)/2, \ 1 + \sum_{t=2}^{n} \epsilon_t^2/2 \right),$$

where $\mathcal{IG}(a, b)$ is the inverse Gamma distribution with shape parameter a and scale parameter b. The full conditional distributions of γ is not explicit, but satisfies

$$P_3(\gamma \mid Y, \beta, \sigma^2) \propto \exp\left\{ -\sum_{t=2}^{n} \epsilon_t^2(\gamma)/(2\sigma^2) - \gamma \right\} \mathbf{1}_{\{\gamma > 0\}}, \quad \epsilon_t(\gamma) = Y_t - \beta' Z_{t-1}.$$

Using a one-dimensional Metropolis-Hastings step to simulate from the conditional distribution P_3, we obtain an hybrid method based on a combination of Metropolis and Gibbs sampling steps.

```
> p3<-function (gam,beta,sig2,y,n) {if ( gam<=0 ) return(-Inf)
+ Zt <- matrix(ncol = n, nrow = 4); Gt<-1/(1+exp(-gam*y))
+ Zt[1,]<- 1-Gt; Zt[2,]<- y*(1-Gt); Zt[3,]<- Gt; Zt[4,]<- y*Gt
+ eps<-c(rep(0,n)); eps[2:n]<-y[2:n]-beta%*%Zt[1:4,1:(n-1)]
+ -gam-sum(eps^2)/(2*sig2) }
>
> gibbsMH<-function (y,nbatch=1000,theta.init=0.1,tun=1) {
+ n<-length(y); eps<-c(rep(0,n)); Zt <- matrix(nrow=4, ncol=n)
+ mat <- matrix(nrow=6, ncol=nbatch); mat[,1] <- theta.init
+  for (i in 2:nbatch) { Gt<-1/(1+exp(-mat[5,i-1]*y))
+   Zt[1,]<- 1-Gt; Zt[2,]<- y*(1-Gt); Zt[3,]<- Gt; Zt[4,]<- y*Gt
+   ZtZ <- Zt[1:4,1:(n-1)]%*%t(Zt[1:4,1:(n-1)])
+   Zty<-{c(sum(y[2:n]*Zt[1,1:(n-1)]),sum(y[2:n]*Zt[2,1:(n-1)]),
+          sum(y[2:n]*Zt[3,1:(n-1)]),sum(y[2:n]*Zt[4,1:(n-1)]))}
+   bhat<-solve(ZtZ,Zty);  Sigmaninv<-ZtZ/mat[6,i-1]
+   Sigma<-solve(diag(rep(1,4))+Sigmaninv)
+   mu<-Sigma%*%(Sigmaninv%*%bhat)
+   mat[1:4,i]<-mvrnorm(n=1,as.vector(mu), Sigma)
+   eps[2:n]<-y[2:n]-bhat%*%Zt[1:4,1:(n-1)]
+   mat[6,i]<- rinvgamma(1,shape=1+(n-1)/2,scale=1+sum(eps^2)/2)
+   gamstar <- rnorm(1, mean=mat[5,i-1], sd=tun)
+   unif<-runif(1,0,1)
+   lognum<-p3(gamstar,bhat,mat[6,i],y,n)
+   logden<-p3(mat[5,i-1],bhat,mat[6,i],y,n)
+   if(lognum==-Inf) ratio<-0 else {if (logden==-Inf)
+            ratio<-1 else ratio<-exp(lognum-logden)}
+   mat[5,i] <- if(unif <= ratio) gamstar else mat[5,i-1]
+    }
+ t(mat)}
>
```

Figure 15 shows the traces and the posterior distributions of γ and σ^2, obtained for a run of length `nbatch=1000` *of the hybrid MCMC algorithm.*

For more information on MCMC methods and on Bayesian statistics the reader is referred to *e.g.* [104].

7.2 Optimization algorithms for models with several latent processes

We have seen in Section 5 that the inference of nonlinear time series models requires the optimization of complicated objective functions of the form (38). For the models considered in Section 5, Newton-type algorithms are sufficient.

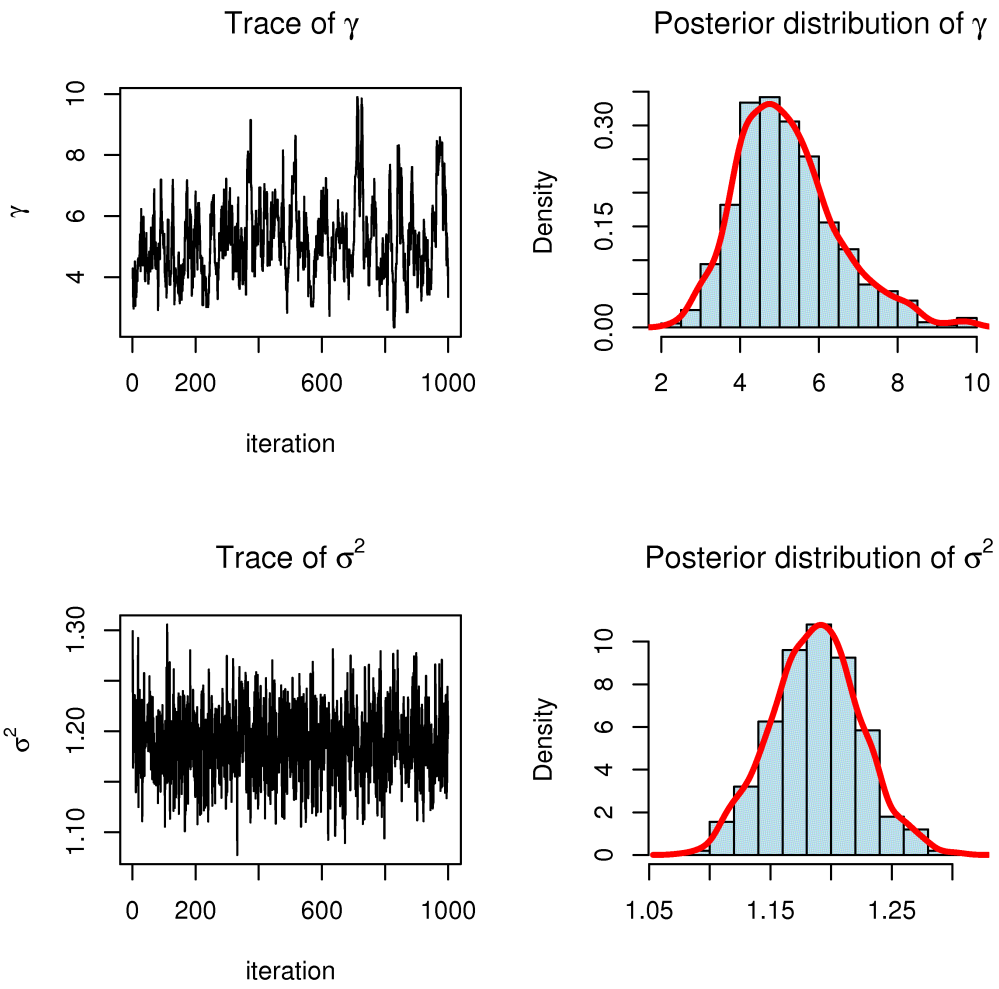


Figure 15. Trace and posterior distribution of the parameters γ and σ^2 of the STAR model (47), obtained with an hybrid Metropolis-Gibbs algorithm.

The problem is much more difficult for models defined by means of several unobservable processes. As an illustration, consider the Markov-switching ARCH model introduced by [72] and defined by

$$\begin{cases} \epsilon_t = \sigma_t \eta_t, \quad \epsilon_t \sim \mathcal{N}(0,1) \\ \sigma_t^2 = \sigma_t^2(\Delta_t) = \omega(\Delta_t) + \sum_{i=1}^{q} \alpha_i(\Delta_t)\epsilon_{t-i}^2 \end{cases} \tag{48}$$

with the same notation as in (4) and (27) for (Δ_t), and the positivity constraints

$$\text{for } j = 1, \ldots, d, \;\; \omega(j) > 0, \;\; \alpha_i(j) \geq 0, \;\; 1 \leq i \leq q.$$

The unknown parameter θ contains the ARCH coefficients of the d regimes and all the non redundant transition probabilities. Assuming that the two latent processes (Δ_t) and (η_t) are independent, conditional on initial values $\epsilon_0, \ldots, \epsilon_{1-q}$, the likelihood of $(\epsilon_1, \ldots, \epsilon_n)$ is given by summing the conditional

density over all the possible paths $(e_1, \ldots, e_n)$ of the Markov chain

$$L(\epsilon_1, \ldots, \epsilon_n; \theta) = \sum_{(e_1,\ldots,e_n)\in\mathcal{E}^n} L^{(e_1,\ldots,e_n)}(\epsilon_1, \ldots, \epsilon_n) P(e_1, \ldots, e_n), \qquad (49)$$

where

$$P(e_1, \ldots, e_n) = P(\Delta_1 = e_1, \ldots, \Delta_n = e_n) = \pi(e_1)p(e_1, e_2)\ldots p(e_{n-1}, e_n)$$

and, denoting by ϕ the $\mathcal{N}(0, 1)$ probability density,

$$L^{(e_1,\ldots,e_n)}(\epsilon_1, \ldots, \epsilon_n) = \prod_{t=1}^{n} \phi_{e_t}(\epsilon_t, \ldots, \epsilon_{t-q}),$$
$$\phi_j(\epsilon_t, \ldots, \epsilon_{t-q}) = \frac{1}{\sigma_t(j)}\phi\left(\frac{\epsilon_t}{\sigma_t(j)}\right).$$

The computation of the likelihood by means of the direct application of (49) is generally impossible because the number of possible paths is d^n, which is a huge number. When $n \geq 300$, the number 2^n is greater than the number of atoms in the universe !

7.2.1 Computation of the likelihood

There exist at least 3 ways to compute the likelihood defined by (49): the forward-backward algorithm defined by [15] (and improved by [48]), the filter defined by [72], and the matrix form that we now present.

Let $F_k(j) = g_k(\epsilon_1, \ldots, \epsilon_k|\Delta_k = j)\pi(j)$ where $g_k(\cdot|\Delta_k = j)$ is the density of $(\epsilon_1, \ldots, \epsilon_k)$ given $\{\Delta_k = j\}$ and the initial values $\epsilon_0, \ldots, \epsilon_{1-q}$. It is easy to check that

$$F_1(j) = \pi(j)\phi_j(\epsilon_1, \ldots, \epsilon_{1-q}) \qquad (50)$$
$$F_k(j) = \phi_j(\epsilon_k, \ldots, \epsilon_{k-q}) \sum_{\ell=1}^{d} F_{k-1}(\ell)p(\ell, j) \qquad (51)$$

and

$$L(\epsilon_1, \ldots, \epsilon_n) = \sum_{j=1}^{d} F_n(j). \qquad (52)$$

In matrix form we obtain

$$F_k := (F_k(1), \ldots, F_k(d))' = M(\epsilon_k, \ldots, \epsilon_{k-q})F_{k-1},$$

where, for $x = (x_1, \ldots, x_{q+1})$,

$$M(x) = \begin{pmatrix} p(1,1)\phi_1(x) & \cdots & p(d,1)\phi_1(x) \\ \vdots & & \\ p(1,d)\phi_d(x) & \cdots & p(d,d)\phi_d(x) \end{pmatrix}.$$

Thus, with $\mathbb{1}' = (1, \ldots, 1)$,

$$L(\epsilon_1, \ldots, \epsilon_n) = \mathbb{1}' M(\epsilon_n, \ldots, \epsilon_{n-q}) \cdots M(\epsilon_2, \ldots, \epsilon_{2-q}) F_1, \tag{53}$$

which is easily computable (with $O(d^2 n)$ multiplications). This matrix form is convenient for studying the asymptotic behavior of the QMLE (see [58]), but is outperformed by the forward-backward algorithm and by Hamilton's filter in terms of computation time, and also in terms of numerical stability. Indeed the matrix product (53) is likely to produce underflows when n is large.

7.2.2 Optimization of the likelihood

The optimization can be performed by means of Newton-type algorithms. The optimization is however subject to inequality constraints: positivity constraints for all the parameters, constraints on the transition probabilities $\sum_{j=1}^{d} p(i,j) = 1$ for $i = 1, \ldots, d$, and the identifiability constraints $\omega(1) \leq \cdots \leq \omega(d)$. Such constraints are easily incorporated in R using the function `constrOptim()`, but the optimization is time consuming. The Expectation-Maximization (EM) algorithm is an interesting alternative which is particularly attractive in the case $q = 0$, because it takes the following explicit form : starting from initial values of the parameters $\pi_0 = \{P(\Delta_1 = 1), \ldots, P(\Delta_1 = d)\}'$, $p(i,j) = P(\Delta_t = j \mid \Delta_{t-1} = i)$ and $\omega = \{\omega(1), \ldots, \omega(d)\}'$,

`repeat the following steps until convergence`

`(1) Set` $\pi_{1|0} = \pi_0$ `and`

$$\pi_{t|t} = \frac{\pi_{t|t-1} \odot \phi(\epsilon_t)}{\mathbb{1}' \left\{ \pi_{t|t-1} \odot \phi(\epsilon_t) \right\}}, \quad \pi_{t+1|t} = \mathbb{P}' \pi_{t|t}, \quad \texttt{for } t = 1, \ldots, n.$$

`(2) Compute the smoothed probabilities` $\pi_{t|n}(i) = P(\Delta_t = i \mid \epsilon_1, \ldots, \epsilon_n)$

$$\pi_{t-1|n}(i) = \sum_{j=1}^{d} \frac{p(i,j)\pi_{t-1|t-1}(i)\pi_{t|n}(j)}{\pi_{t|t-1}(j)} \quad \texttt{for } t = n, n-1, \ldots, 2.$$

`and the probabilities` $\pi_{t-1,t|n}(i,j) = P(\Delta_{t-1} = i, \Delta_t = j \mid \epsilon_1, \ldots, \epsilon_n)$ `from`

$$\pi_{t-1,t|n}(i,j) = \frac{p(i,j)\pi_{t-1|t-1}(i)\pi_{t|n}(j)}{\pi_{t|t-1}(j)}.$$

(3) Replace the previous values of the parameters by $\pi_0 = \pi_{1|n}$,

$$p(i,j) = \frac{\sum_{t=2}^{n} \pi_{t-1,t|n}(i,j)}{\sum_{t=2}^{n} \pi_{t-1|n}(i)} \text{ and } \omega(i) = \frac{\sum_{t=1}^{n} \epsilon_t^2 \pi_{t|n}(i)}{\sum_{t=1}^{n} \pi_{t|n}(i)}.$$

In (1)-(3) the symbol $\odot$ denotes the Adamar product, $\mathbb{P}$ is the matrix of the transition probabilities, and $\phi(\epsilon_t) = \{\phi_1(\epsilon_t), \ldots, \phi_d(\epsilon_t)\}'$.

The reader is referred to [73] and [82] for details on this algorithm and its extensions. In the HMM literature, the part (1) is generally replaced by the Forward-Backward algorithm, and (2) is then obtained by the Viterbi algorithm, whereas (3) is known as the Baum-Welch algorithm (see *e.g.* [103]).

Example 19 *The considered the French CAC 40 stock index from 1 Mars 1990 to 29 December 2006. On the daily returns (in %), we fitted the Markov-switching model (48) with $q = 0$ and $d = 4$ regimes, using the EM algorithm. The estimated parameters are*

$$\hat{\omega} = \begin{pmatrix} 0.51 \\ 1.19 \\ 2.45 \\ 8.4 \end{pmatrix}, \quad \hat{\mathbb{P}} = \begin{pmatrix} 0.993 & 0.003 & 0.002 & 0.002 \\ 0.003 & 0.991 & 0.003 & 0.003 \\ 0.000 & 0.020 & 0.977 & 0.003 \\ 0.004 & 0.000 & 0.032 & 0.963 \end{pmatrix}.$$

Note that the estimated probabilities of the regimes are $\hat{\pi} = (0.26, 0.49, 0.19, 0.06)'$, and that the average duration of the regimes are respectively $140, 107, 43,$ *and* 27 *days ($1/(1-p(i,i))$ for $i = 1, \ldots, 4$). Figure16 confirms that the regime with the highest volatility is the less frequent and the less persistent, with however an exceptional long period of high volatility from 4 June 2002 to 8 November 2002. The regime with the lowest volatility is the most persistent, and the second regime is the most frequent.*

8 Conclusion

Most of the real-life time series, in particular those encountered in finance and macroeconomics, exhibit nonlinearities. Conventional time series models, like the ARMA processes, are inappropriate for such series. This is the reason why, in econometrics and statistics, the literature on nonlinear time series models has developed considerably in the last decade. The present work aims to give an idea of the variety of the methods that are employed in nonlinear time series analysis. The themes tackled in this chapter have been chosen to illustrate that nonlinear time series analysis requires the interaction of probability theory, statistical inference, applied econometrics and computational

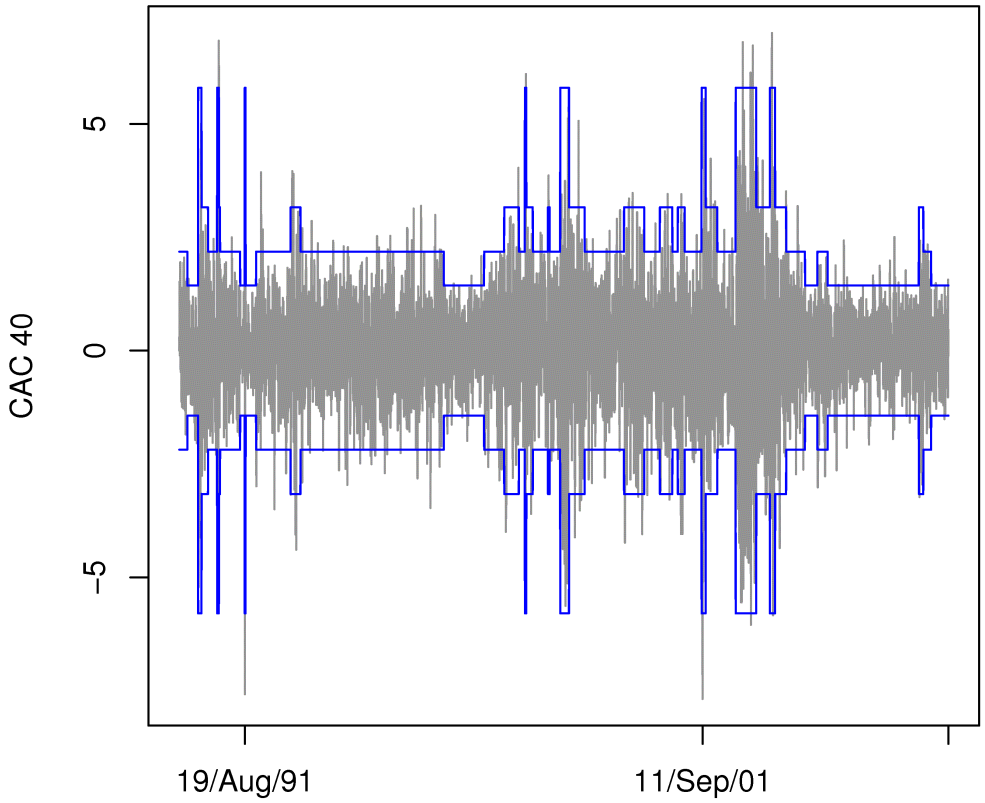


Figure 16. Returns of the CAC 40 index from 2 Mars 1990 to 29 December 2006, and ±2 times the standard deviation of the regime which maximizes the smoothed probability.

methods. The paper concentrates on a selection of application-oriented tools and concepts which covers all the above-mentioned domains. Original examples and illustrations are given throughout the text. Most of the examples are deliberately simplistic, the goal being to illustrate the main ideas. An application to the regime changes in the volatility of the CAC 40 stock index is also given. It should be emphasized, however, that in real applications, the multiplicity of the stylized facts leads to the introduction of highly complex models for which the modelling issues are far from being obvious. This is why nonlinear time series modelling is a very active area of academic research, with numerous exciting problems.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in Second International Symposium in Information Theory, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest (2003) 267–281.

[2] M.S. Al-Qassem, J.A. Lane, Forecasting exponential autoregressive models of

order 1, Journal of Time Series Analysis, 10 (1989) 95–113.

[3] A. Amendola, M. Niglio, Predictor distribution and forecast accuracy of threshold models, Statistical Methods & Applications, 13 (2003) 3–14.

[4] A. Amendola, M. Niglio, C. Vitale, The moments of SETARMA models, Journal of Statistics & Probability Letters, 76 (2006) 625–633.

[5] A. Amendola, M. Niglio, C. Vitale, The multi-step ahead predictors of SETARMA models, mimeo (2006).

[6] T.G. Andersen, T. Bollerslev, F.X. Diebold, P. Labys, Modeling and Forecasting Realized Volatility, Econometrica, 71 (2005) 579–625.

[7] T.G. Andersen, T. Bollerslev, Volatility and Correlation Forecasting, Handbook of Economic Forecasting, ed. C.W.J. Granger and A. Timmermann, Elsevier, Amsterdam, (2006).

[8] D.W.K. Andrews, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, Econometrica, 59 (1991) 817–858.

[9] D.W.K. Andrews, W. Ploberger, Optimal tests when a nuisance parameter is present only under the alternative, Econometrica, 62 (1994) 1383–1414.

[10] J. Bai, Testing Parametric Conditional Distributions of Dynamic Models, Review of Economics and Statistics, 85 (2003) 531–549.

[11] R.T. Baillie, T. Bollerslev, Prediction in dynamic models, Journal of Econometrics, 52 (1992) 91–113.

[12] Y. Bao, T.H. Lee, B. Saltoglu, Comparing Density Forecast Models, Journal of Forecasting, 26 (2007) 203–225.

[13] I.V. Basawa, R. Lund, Large sample properties of parameter estimates for periodic ARMA models, Journal of Time Series Analysis, 22 (2001) 651–663.

[14] J.M. Bates, C.W.J. Granger, The combination of forecasts, Operational Research Quarterly, 30 (1969) 451–468.

[15] L.E. Baum, An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, Inequalities, 3 (1972) 1–8.

[16] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Annals of Mathematical statistics, 30 (1966) 1554–1563.

[17] J. Berkowitz, Testing Density Forecasts with Applications to Risk Management, Journal of Business and Economic Statistics, 19 (2001) 465–474.

[18] A. Bibi, A. Gautier, Propriétés dans L^2 et estimation des processus purement bilinéaires et strictement superdiagonaux à coefficients périodiques, La revue Canadienne de Statistique, (2006) to appear.

[19] P. Billingsley, The Lindeberg-Levy theorem for martingales, Proceedings of the American Mathematical Society, 12 (1961) 788–792.

[20] P. Billingsley, Probability and Measure, Wiley, New-York, (1995).

[21] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics, 31 (1986) 307–327.

[22] P. Bougerol, N. Picard, Strict stationarity of generalized autoregressive processes, Annals of Probability, 20 (1992a) 1714–1729.

[23] P. Bougerol, N. Picard, Stationarity of GARCH processes and of some nonnegative time series, Journal of Econometrics, 52 (1992b) 115–127.

[24] G.E.P. Box, G.M. Jenkins, Time series analysis, forecasting and control, San Francisco, Holden-Day (1976).

[25] G.E.P. Box, D.A. Pierce, Distribution of Residual Autocorrelations in Autoregressive-Integreted Moving Average Time Series Models, Journal of the American Statististical Association, 65 (1970) 1509–1526.

[26] A. Brandt, The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients, Advances in Applied Probability, 18 (1986) 221–254.

[27] F.J. Breidt, R.A. Davis, Time-reversibility, identifiability, and independence of innovations for stationary time series, Journal of Time Series Analysis, 13 (1992) 377–390.

[28] F.J. Breidt, R.A. Davis, A. Trindade, Least Absolute Deviation Estimation for All-Pass Time Series Models, Annals of Statistics, 29 (2001) 919–946.

[29] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, Springer-Verlag, New-York (1991).

[30] S.P. Brooks, G.O. Roberts, Assessing Convergence of Markov Chain Monte Carlo Algorithms, Statistics and Computing, 8 (1998) 319–335.

[31] B.Y. Brown, R.S. Mariano, Predictors dynamic nonlinear models: large sample behaviour, Econometric Theory, 5 (1989) 430–452.

[32] C. Chatfield, Calculating interval forecasts, Journal of Business and Economic Statistics, 11 (1993) 121–144.

[33] R. Chen, R.S. Tsay, Functional-Coefficient Autoregressive Models, Journal of the American Statistical Association, 88 (1993) 298–308.

[34] Q. Cheng, On time-reservibility of linear processes, Biometrika, 86 (1999) 483–486.

[35] P. Christoffersen, Evaluating Interval Forecasts, International Economic Review, 39 (1998) 841–62.

[36] P. Christoffersen, F.X. Diebold, How Relevant is Volatility Forecasting for Financial Risk Management?, Review of Economics and Statistics, 82 (2000) 12–22.

[37] M.P. Clements, J. Smith, The performance of alternative forecasting methods for SETAR models, International Journal of Forecasting, 13 (1997) 463–475.

[38] M.P. Clements, P.H. Franses, L. Smith, D. Van Dijk, On SETAR non-linearity and forecasting, Journal of Forecasting, 22 (2003) 359–375.

[39] M.P. Clements, P.H. Franses, N.R. Swanson, Forecasting economic and financial time-series with non-linear models, International Journal of Forecasting, 20 (2004) 169–183.

[40] D.B.H. Cline, Evaluating the Lyapounov exponent and existence of moments for threshold AR-ARCH models, Preprint (2006).

[41] D.B.H. Cline, H.H. Pu, Stability and the Lyapounov exponent of threshold AR-ARCH models, The Annals of Applied Probability, 14 (2004) 1920–1949.

[42] V. Corradi, N.R. Swanson, Predictive Density Evaluation, Handbook of Economic Forecasting, Vol. 1, ed. Elliott G., Granger C. and Timmermann A., Amsterdam: North-Holland, (2006).

[43] J. Davidson, Stochastic Limit Theory, Oxford University Press, Oxford (1994).

[44] J. Davidson, Stochastic Limit Theory, Oxford University Press, Oxford (1994).

[45] R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika, (1977) 64 247–254.

[46] R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika, 74 (1987) 33-43.

[47] Y.A. Davydov, Convergence of Distributions Generated by Stationary Stochastic Processes, Theory of Probability and Its Applications, 13 (1968) 691–696.

[48] P. Devijver, Baum's forward-backward algorithm revisited, Pattern Recognition Letters, 3 (1985) 369–373.

[49] F.X. Diebold, T.G. Gunther, A.S. Tay, Evaluating Density forecast with applications to Financial Risk Management, International Economic Review, 39 (1998) 863–883.

[50] Z. Ding, C. Granger, R.F. Engle, A long memory property of stock market returns and a new model, Journal of Empirical Finance, 1 (1993) 83–106.

[51] R.F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation, Econometrica, 50 (1982) 987–1007.

[52] R.F. Engle, Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, Handbook of Econometrics, Vol. 2, ed. Z. Griliches and M.D. Intriligator, Amsterdam: North-Holland, (1984) 775–826.

[53] R.F. Engle, D.M. Lilien, R.P. Robins, Estimating time varying riak premia in the term structure: the ARCH-M mode, Econometrica, 55 (1987) 391–407.

[54] R.F. Engle, V.K. Ng, Measuring and Testing the Impact of News on Volatility, The Journal of Finance, 48 (1993) 1749–1778.

[55] J. Fan, Q. Yao, Nonlinear Time Series : Nonparametric and Parametric Methods. Springer-Verlag, New York (2003).

[56] P.D. Feigin, R.L. Tweedie, Random Coefficient Autoregressive Processes: a Markov Chain Analysis of Stationarity and Finiteness of Moments, Journal of Time Series Analysis, 6 (1985) 1–14.

[57] D.F. Findley, The overfitting principles supporting AIC, Statistical Research Division Report RR 93/04, Bureau of the Census (1993).

[58] C. Francq, M. Roussignol, J.M. Zakoïan, Conditional heteroskedasticity driven by hidden Markov chains. Journal of Time Series Analysis, 22 (2001) 197–220.

[59] C. Francq, R. Roy, J.M. Zakoïan, Diagnostic Checking in ARMA Models with Uncorrelated Errors, Journal of the American Statistical Association, 100 (2005) 532–544.

[60] C. Francq, J.M. Zakoïan, Stationarity of Multivariate Markov-switching ARMA Models, Journal of Econometrics, 102 (2001) 339–364.

[61] C. Francq, J.M. Zakoïan, Recent results for linear time series models with non independent innovations, in Statistical Modeling and Analysis for Complex Data Problems, Duchesne, P. et Rémillard, B., Éditeurs, Kluwer, (2004) 241–266.

[62] P.H. Franses, D. van Dijk, Non-linear time series Models in empirical finance, Cambrige University Press, Cambrige (2002).

[63] S. Geman, D. Geman, Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (1984) 721–741.

[64] L.R. Glosten, R. Jagannathan, D. Runkle, On the relation between the expected value and the volatility of the nominal excess return on stocks, Journal of Finance, 48 (1993) 1779–1801.

[65] L.G. Godfrey, Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches, Cambridge University Press, Cambridge (1988).

[66] J.B. de Gooijer, P.T. de Bruin, On forecasting SETAR processes, Statistics & Probability Letters, 37 (1998) 7–14.

[67] C.W.J. Granger, A.P. Andersen, An introduction to bilinear time series models. Vandenhoek and Ruprecht, Gottingen (1978).

[68] C.W.J. Granger, R. Ramanathan, Improved methods of combining forecasts, Journal of Forecasting, 3 (1984) 197–204.

[69] C.W.J. Granger, T. Teräsvirta, Modelling nonlinear relationships, Oxford University Press, Oxford (1993).

[70] V. Haggan, T. Ozaki, Modelling Nonlinear Random Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model, Biometrika, 68 (1981) 189–196.

[71] S.G. Halla, J. Mitchell, Combining density forecasts, International Journal of Forecasting, 23 (2007) 1–13.

[72] J.D. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica, 57 (1989) 357–384.

[73] J.D. Hamilton, Time Series Analysis, Princeton University Press ed, (1994).

[74] B.E. Hansen, Inference when a nuisance parameter is not identified under the null hypothesis, Econometrica, 64 (1996) 413-430.

[75] W. Härdle, Applied Nonparametric Regression, Cambridge University Press: New York (1990).

[76] W.K. Hastings, Monte Carlo Sampling Methods using Markov Chains and their Applications, Biometrika, 57 (1970) 97–109.

[77] N. Herrndorf, A Functional Central Limit Theorem for Weakly Dependent Sequences of Random Variables, The Annals of Probability, 12 (1984) 141–153.

[78] Y. Hong, Evaluation of out-of-sample Probability Density Forecasts with Applications to S&P 500 Stock Prices, Working Paper, (2001) Cornell University.

[79] Y. Hong, Y.J. Lee, Consistent Testing for Serial Correlation of Unknown Form Under General Conditional Heteroskedasticity, Preprint, Cornell University, (2003).

[80] R.J. Hyndman, Highest-density forecast regions for non-linear and non-normal time series models, Journal of Forecasting, 14 (1995) 431–441.

[81] R.J. Hyndman, Computing and graphing highest density regions, American Statisticians, 50 (1996) 120–126.

[82] H.K. Krolzig, Markov Switching Vector Autoregressions. Modelling Statistical Inference and Application to Business Cycle Analysis, Springer Verlag ed (1997).

[83] W.K. Li, Diagnostic Checks in Time Series. Chapman & Hall/CRC, Boca Raton, Florida (2004).

[84] W.K. Li, K. Lam, Modelling asymmetry in stock returns by threshold autoregressive conditional heteroscedastic model, The Statistician, 44 (1995) 333–341.

[85] C.W. Li, W.K. Li, On a double-threshold heteroscedastic time series model, Journal Applied Econometric, 11 (1996) 253–274.

[86] J.L. Lin, C.W.J. Granger, Forecasting from non-linear models in practice, Journal of Forecasting, 13 (1994) 1–19.

[87] J. Liu, W.K. Li, C.W. Li, On a threshold autoregression with conditional heteroscedastic variances, Journal of Statistical Planning and inference, 62 (1997) 279–300.

[88] G.M. Ljung, G.E.P. Box, On the Measure of Lack of Fit in Time Series Models, Biometrika, 65 (1978) 297–303.

[89] I.N. Lobato, Testing for Zero Autocorrelation in the Presence of Statistical Dependence, Econometric Theory, 18 (2002) 730–743.

[90] R. Luukkonen, P. Saikkonen, T. Teräsverta, Testing linearity against smooth transition autoregression, Biometrika, 75 (1988) 491–499.

[91] A.I. McLeod, W.K. Li, Diagnostic checking ARMA time series models using squaredresidual autocorrelations, Journal of Time Series Analysis, 4 (1983) 269–273.

[92] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Application in Statistics and Econometrics, New-York, Wiley (1988).

[93] M. Marcellino, Forecasting EMU macroeconomic variables, International Journal Of Forecasting, 20 (2004) 359–372.

[94] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equations of State Calculations by Fast Computing Machines, Journal of Chemical Physics, 21 (1953) 1087–1092.

[95] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic stability, 3rd Edition, Springer, London (1996).

[96] R. Moenaddin, Approximating multi-step non-linear least squares prediction of threshold autoregressive models, Paper presented at the IMS Philadelphia meeting, (1991).

[97] D.B. Nelson, Conditional Heteroskedasticity in Asset Returns : a New Approach, Econometrica, 59 (1991) 347–370.

[98] D.F. Nicholls, B.G. Quinn, Random Coefficient Autoregressive Models: An Introduction. Lecture Notes in Statist. Springer-Verlag, New York (1982).

[99] A. Patton, Volatility Forecast Comparison using Imperfect Volatility Proxies, Research Paper Series 175, Quantitative Finance Research Centre, University of Technology, Sydney (2006).

[100] B.M. Pötscher, I.R. Prucha, Dynamic Nonlinear Econometric Models, Springer, Berlin (1997).

[101] S.M. Potter, Nonlinear time series modelling: an introduction. Journal of Economic Survey, 13 (1999) 505–528.

[102] M.B. Priestley, Nonlinear and non-stationary time series analysis, Academic press, London, 1988.

[103] L. Rabiner, B. Juang, An introduction to hidden Markov models, ASSP Magazine, IEEE, 3 (1986) 4–16.

[104] C.P. Robert, G. Casella, Monte Carlo Statistical Methods. 2nd Edition. New York: Springer (2004).

[105] J.P. Romano, L.A. Thombs, Inference for autocorrelations under weak assumptions, Journal of the American Statistical Association, 91 (1996) 590–600.

[106] E. Rio, Covariance inequalities for strongly mixing processes, Ann. Inst. Henri Poincaré, 29 (1993) 587–597.

[107] E. Sentana, Quadratic ARCH Models, Review of Economic Studies, 62 (1995) 639–661.

[108] J.H. Stock, M.W. Watson, A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series, in R.F. Engle and h. White (eds), Cointegration, Causality and Forecasting. A Festschrift in Honour of Clive W.J. Granger, Oxford University Press, Oxford, (1999) 1–44.

[109] T. Subba Rao, M.M. Gabr, An introduction to bispectral analysis and bilinear time series models. Lecture Notes in Statistics, Springer, New-York (1984).

[110] S.J. Taylor, Modelling Financial Time Series, New-York: Wiley (1986).

[111] T. Teräsvirta, Smooth Transition Regression Modeling, in Applied Time Series Econometrics Eds. H. Lütkepohl and M. Krätzig, Cambridge University Press, Cambridge (2004) 222–242.

[112] T. Teräsvirta, Forecasting economic variables with nonlinear models, Handbook of Economic Forecasting, ed. G. Elliott, C.W.J. Granger and A. Timmermann, Elsevier, Amsterdam, (2006).

[113] T. Teräsvirta, D. Tjøstheim, C.W.J. Granger, Aspects of Modelling Nonlinear Time Series, Handbook of Econometrics, Vol. 4, ed. R.F. Engle and D.L. McFadden, Elsevier, Amsterdam, (1994) 2919–2960.

[114] A. Timmermann, Forecast Combination, Handbook of Economic Forecasting, Vol.1, ed. Elliott G., Granger C. and Timmermann A., Amsterdam: North-Holland, (2006) 135–194.

[115] D. Tjøstheim, Estimation in nonlinear time series models, Stochastic Processes and their Applications, 21 (1986) 251–273.

[116] D. Tjøstheim, Non-linear time series and Markov chains, Advances in Applied Probability, 22 (1990) 587–611.

[117] D. Tjøstheim, B.H. Auestad, Nonparametric identification of nonlinear time series: projections, Journal of the American Statistical Association, 89 (1994) 1398–1409.

[118] H. Tong, A note on using threshold autoregressive models for multi-step-ahead prediction of cyclical data, Journal of the Time Series Analysis, 3 (1982) 137–140.

[119] H. Tong, Threshold models in nonlinear time series analysis, Springer-Verlag, New York (1983).

[120] H. Tong, Non-Linear Time Series. A dynamical System Approach, Oxford University Press, Oxford (1990).

[121] H. Tong, K.S. Lim, Threshold autoregression, limit cycles and cyclical data (with discussion), Journal of the Royal Statististical Society, Series B, 42 (1980) 245–292.

[122] H. Tong, R. Moennaddin, On multi-step non-linear least squares prediction, The Statistician, 37 (1988) 101–110.

[123] R.S. Tsay, Analysis of Financial Time Series, Wiley 2nd Edition (2005).

[124] K.F. Wallis, Asymmetric density forecasts of inflation and the Bank of England's fan chart, National Institute Economic Review, 167 (1999) 106–112.

[125] K.F. Wallis, Combining Density and Interval Forecasts: A modest proposal, Oxford Bulletin of Economics and Statistics, 67 (2005) 983–994.

[126] H. Wold, A study in the analysis of stationary time series, Almqvist and Wiksell, Stocholm (1938).

[127] Q. Yao, H. Tong, Quantifying the influence of initial values on nonlinear prediction, Journal of the Royal Statistical Society, Series B, 56 (1994) 701–725.

[128] J.M. Zakoïan, Threshold Heteroskedastic Models, Journal of Economic Dynamics and Control, 18 (1994) 931–955.