Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 161 (2010) 1836–1851
FUZZY sets and systems
www.elsevier.com/locate/fss

Linearity testing for fuzzy rule-based models

José Luis Aznarte[a,*,1], Marcelo C. Medeiros[b], José M. Benítez[c]

[a]*Department of Computer Science and Artificial Intelligence, CITIC-UGR, University of Granada, Spain*
[b]*Department of Economics, Pontifical Catholic University of Rio de Janeiro, Brazil*
[c]*Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communication Technology), University of Granada, Spain*

Received 24 March 2009; received in revised form 31 December 2009; accepted 30 January 2010
Available online 2 February 2010

Abstract

In this paper, we introduce a linearity test for fuzzy rule-based models in the framework of time series modeling. To do so, we explore a family of statistical models, the regime switching autoregressive models, and the relations that link them to the fuzzy rule-based models. From these relations, we derive a Lagrange multiplier linearity test and some properties of the maximum likelihood estimator needed for it. Finally, an empirical study of the goodness of the test is presented.

Keywords: Fuzzy rule-based models; Time series; Linearity test; Statistical inference

1. Introduction

Time series analysis is a problem which has always attracted the attention of soft computing (SC) researchers. Forecasting future values of a series is usually a very complex task, and many SC methods and models have been faced with it, including fuzzy rule-based models (FRBM) in their various formulations [18]. Notwithstanding, a common characteristic of those approaches is that they usually consider time series as just another dataset which requires some small adaptations to be cast into the regression or classification format for which most SC models were created.

However, time series analysis is a prominent field in Econometrics, which has been widely studied under a statistical perspective during the last centuries. In 1807, Fourier proved that a deterministic time series can be approximated by a sum of sine and cosine terms. But it was not until the beginnings of the 20th century when a stochastic approach for time series was first introduced [4], while the foundations for a general stochastic process theory were fixed in the 1930s by [13,12]. Independently, in 1927 Yule [31] stated that Fourier analysis is not suited for stochastic time series analysis and introduced second order autoregressive processes as theoretical schemes able to generate series with stochastic cyclic oscillations.

In 1970, the idea of forecasting future values of a time series as a combination of its past values received a strong impulse after [6]. In that work, Box and Jenkins proposed a modeling cycle for the autoregressive (AR) model, which assumes that future values of a time series can be expressed as a linear combination of its past values.

* Corresponding author. Tel.: +34 958246143; fax: +34 958 243317.
E-mail address: jlaznarte@decsai.ugr.es (J.L. Aznarte).
[1] Now with the Centre for Energy and Processes, MINES ParisTech, France.

0165-0114/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2010.01.005

Of course this linearity assumption implies certain limitations, and in the last years much research has been devoted to nonlinear models. Nonlinear and non-stationary models are more flexible in capturing the characteristics of data and, in some cases, are better in terms of estimation and forecasting. These advances do not rule out linear models at all, because these models are a first approach which can be of great help to further estimate some of the parameters. Furthermore, modeling of any real-world problem by using nonlinear models must start by evaluating if the behavior of the series follows a linear or nonlinear pattern.

For some reason, SC researchers do not usually go deep into classical time series analysis, disregarding all the knowledge gathered through the years in the statistical field. In this paper, we take a step forward in the quest for an SC-based time series research which integrates methods and models coming from the econometric perspective, introducing a linearity test against fuzzy rule-based models.

By applying this test, practitioners will be able to determine if a series' data generating process is linear, in which case it can be modeled by using a linear model or a single-rule fuzzy rule-based model. The experiments show that the test is robust against Type I errors (rejecting the null hypothesis when it is actually true) and very powerful against Type II errors (not rejecting the null hypothesis when it is false).

The structure of the paper is as follows: in Section 2, a brief review of some statistical regime-switching models is offered, while in Section 3 their links with fuzzy rule-based models are recalled. In Section 4 the linearity test is presented, both intuitively and in its mathematical formulation. Section 5 contains a Monte Carlo simulation and a power analysis of the test, which shows its robustness. Finally, the conclusions of the paper are gathered in Section 6, while Appendix A contains some required results about the properties of the maximum likelihood estimator.

2. Regime switching autoregressive models

In statistical time series modeling, one of the oldest and most successful concepts is to forecast future values of a time series as a combination of its past values. This is a quite natural idea that we apply on every day's life, and it was popularized in 1970 after [6]. In that work, Box and Jenkins formalized the use of the *autoregressive* (AR) model, which assumes that future values of a time series can be expressed as a linear combination of its past values.

An *autoregressive model* of order $p \geq 1$ is defined as

$$y_t = \mathbf{b}'\mathbf{x}_t = b_0 + b_1 y_{t-1} + \cdots + b_p y_{t-p} + \varepsilon_t, \tag{1}$$

where $\mathbf{x}_t = (1, y_{t-1}, \ldots, y_{t-p})'$ and $\{\varepsilon_t\} \sim N(0, \sigma^2)$, usually known as *white noise* (equivalent to a random signal with a flat power spectral density). For this model we write $\{y_t\} \sim \mathrm{AR}(p)$, and the time series $\{y_t\}$ generated from this model is called the $\mathrm{AR}(p)$ process.

Model (1) represents the current state y_t through its immediate p past values $y_{t-1}, \ldots, y_{t-p}$ in a linear regression form. It explicitly specifies the relationship between the current value and its past values.

Such a simple idea proved to be extremely useful and suited to series which, at first sight, seem to be highly complex. Applications of the Box and Jenkins methodology spread in the following decades, covering a wide range of scientific areas such as Biology, Astronomy or Econometrics.

However, there were still many problems which could not be modeled using linear models. In 1978, taking a step towards integrating nonlinearity, Tong [23] proposed a *piecewise linear* model: the threshold autoregressive (TAR) model, which is based on the idea of partitioning the state-space into several subspaces, each of which was to be modeled by an AR model. To control the transitions from one linear model to another, a set of thresholds must be defined on one of the variables involved. This variable can be an exogenous variable associated to the process being modeled or one of the lagged values of the series, in which case the model is called self-exciting—yielding the acronym SETAR.

A *threshold autoregressive* (TAR) model with k ($k \geq 2$) regimes is defined as

$$y_t = \sum_{i=1}^{k} \mathbf{b}_i'\mathbf{x}_t I(s_t \in A_i) + \varepsilon_t = \sum_{i=1}^{k} \{b_{i,0} + b_{i,1} y_{t-1} + b_{i,p} y_{t-p} + \varepsilon_t\} I(s_t \in A_i) + \varepsilon_t, \tag{2}$$

where s_t is the threshold variable, I is an indicator (or *step*) function, $\mathbf{b}_i$ are unknown parameters, and $\{A_i\}$ forms a partition of $(-\infty, \infty)$ with $\bigcup_{i=1}^{k} A_i = (-\infty, \infty)$ and $A_i \cap A_j = \emptyset, \forall i \neq j$.

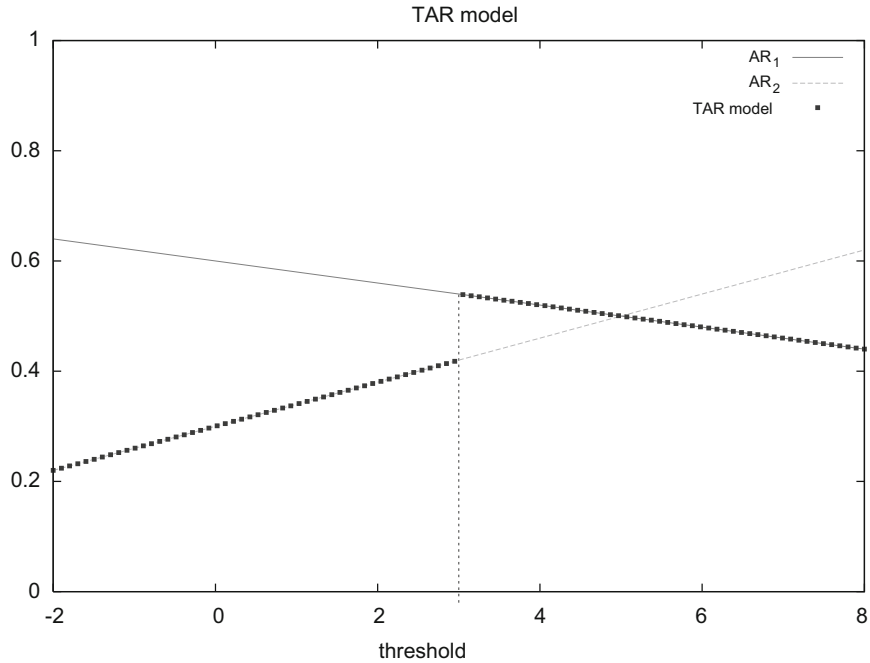

Fig. 1. An example of TAR model.

In this model, we fit on each subset A_i a linear autoregressive form. The partition is dictated by the threshold variable s_t. Usually, $A_i = (r_{i-1}, r_i]$, with $-\infty = r_0 < r_1 < \cdots < r_k = \infty$, where the r_i's are called thresholds.

In Fig. 1 we can see the graphical representation of an example TAR model with two regimes.

2.1. Smooth transition autoregressive model (STAR)

A key feature of TAR models is the discontinuous nature of the AR relationship as the threshold is passed. Taking into account that nature is generally continuous, in 1994 an alternative model called *smooth threshold autoregressive* or *smooth transition autoregressive* (STAR) was proposed by Teräsvirta [21]. In STAR models there is a smooth continuous transition from one linear AR to another, rather than a sudden jump.

In this model and variants (cf. [24]), the indicator function $I(\cdot)$ in (2), which, as shown above, is a *step* function that takes the value zero below the threshold and one above it is substituted by a smooth function with sigmoid characteristics. The STAR model with k+1 regimes is defined as

$$y_t = \mathbf{b}_0' \mathbf{x}_t + \sum_{i=1}^{k} \mathbf{b}_i' \mathbf{x}_t \, f_i(s_t; \boldsymbol{\phi}_i) + \varepsilon_t. \tag{3}$$

The transition function, $f_i(s_t; \boldsymbol{\phi}_i)$, is a continuous function that is bounded between 0 and 1. The regime that occurs at time t is determined by the observable variable s_t and the associated value of $f_i(s_t; \boldsymbol{\phi}_i)$. Different choices for the transition function give rise to different types of regime-switching behavior. A popular choice for $f_i(s_t; \boldsymbol{\phi}_i)$ is the first-order logistic function,

$$f_l(s_t; c, \gamma) = (1 + \exp(\gamma(s_t - c)))^{-1}, \tag{4}$$

and the resultant model is called the logistic STAR (LSTAR).

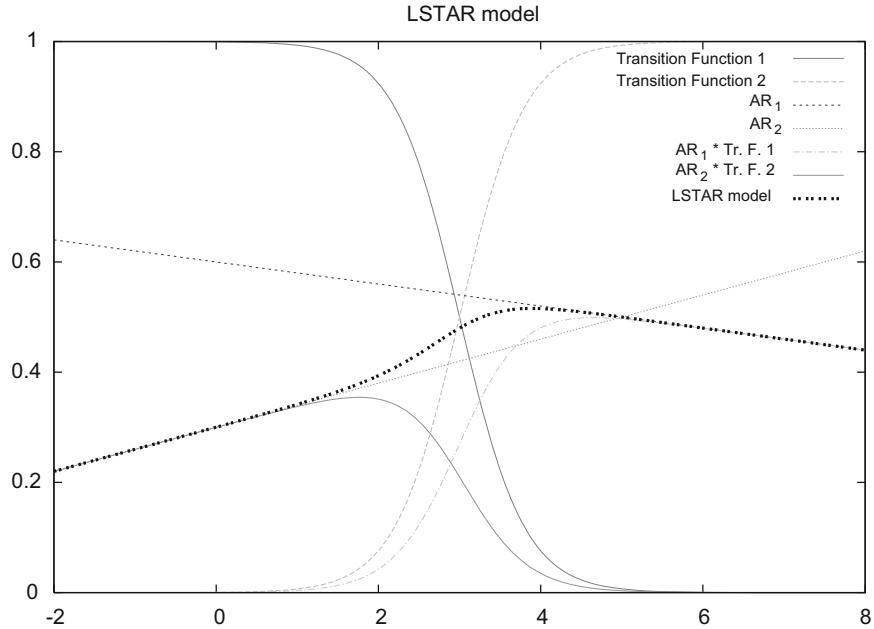

Fig. 2. An example of 2 regime STAR model using logistic transition function.

In the LSTAR model, we define the transition function $\mathrm{F}(s_t; \boldsymbol{\phi}_i)$ of expression (3) as

$$\mathrm{F}_i(s_t; \gamma_i, c_i) = \begin{cases} 1 - f_l(s_t; \gamma_i, c_i) & \text{if } i = 1, \\ f_l(s_t; \gamma_i, c_i) - f_l(s_t; \gamma_{i+1}, c_{i+1}) & \text{if } 1 < i < k, \\ f_l(s_t; \gamma_i, c_i) & \text{if } i = k, \end{cases} \tag{5}$$

where $f_l(s_t; \gamma_i, c_i)$ is defined as in (4). The LSTAR model can be (and usually is) consequently rewritten as

$$y_t = \sum_{i=2}^{k} \mathbf{b}_i' \mathbf{x}_t \mathrm{F}(s_t; \gamma_i, c_i) + \varepsilon_t. \tag{6}$$

Fig. 2 shows the graphical representation of a STAR model with two regimes and two transition functions, together with the linear models and the transition functions associated with them.

Each of the parameters c_i in (6) can be interpreted as the threshold between two regimes, in the sense that the logistic function changes monotonically from 0 to 1 as s_t increases and $\mathrm{F}(c_i; \gamma_i, c_i) = 0.5$. The parameter γ_i determines the smoothness of the transition from one regime to another. As γ_i becomes very large, the logistic function approaches the indicator function $I(\cdot)$ and hence the change of $\mathrm{F}(s_t; \gamma_i, c_i)$ from 0 to 1 becomes instantaneous at $s_t \to c$. Consequently, the LSTAR nests threshold autoregressive (TAR) models as a special case. Furthermore, when $\gamma \to 0$ the LSTAR model reduces to a linear AR model.

In the LSTAR model, the regime switches are associated with small and large values of the transition variable s_t relative to c. In certain applications it may be more appropriate to specify the transition function such that the regimes are associated with small and large absolute values of s_t (again relative to c). This can be achieved by using, for example, the exponential function, in which case the model may be named ESTAR. Other frequently used function is the normal distribution, which yields the acronym NSTAR.

3. Relations with fuzzy rule-based models

In [1] we explored the existing links between an AR model and a fuzzy rule used in the time series framework. As well, we proved that STAR models can be seen as a particular case of a fuzzy rule-based model. Here we will briefly

recall those results. For the sake of clarity, let us first note the expression of the fuzzy rule-based model considered here.

When dealing with time series problems (and, in general, when dealing with any problem for which precision is more important than interpretability), the Takagi–Sugeno–Kang paradigm is preferred over other variants of FRBMs. A fuzzy rule of type TSK has the following shape:

$$\begin{aligned}&\text{IF } x_1 \text{ IS } A_1 \text{ AND } x_2 \text{ IS } A_2 \text{ AND, } \ldots, \text{ AND } x_p \text{ IS } A_p\\ &\quad \text{THEN } y = \mathbf{b}'\mathbf{x}_t = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p,\end{aligned} \tag{7}$$

where x_i are input variables and A_j are fuzzy sets for input variables.

Concerning the fuzzy reasoning mechanism for TSK rules, the *firing strength* of the ith rule is obtained as the t-norm (usually, multiplication operator) of the membership values of the premise part terms of the linguistic variables:

$$\omega(\mathbf{x}) = \prod_{j=1}^{p} \mu_{A_j}(x_j), \tag{8}$$

where the shape of the membership function of the linguistic terms μ_{A_j} can be chosen from a wide range of functions. One of the most common is the Gaussian bell,

$$\mu_A(x) = \exp \frac{-(x-c)^2}{2\sigma^2}, \tag{9}$$

but it can also be a logistic function,

$$\mu_A(x) = \frac{1}{1 + \exp\left(\dfrac{c-x}{\sigma^2}\right)}, \tag{10}$$

and also non-derivable functions as a triangular or trapezoidal function.

The overall output is computed as a weighted average or weighted sum of the rules output. In the case of the weighted sum, the output expression is

$$y_t = G(\mathbf{x}_t; \boldsymbol{\psi}) = \sum_{i=1}^{R} \mathbf{b}_i' \mathbf{x}_t \cdot \omega_i(\mathbf{x}_t), \tag{11}$$

where G is the general nonlinear function with parameters $\boldsymbol{\psi}$, and R denotes the number of fuzzy rules included in the system. While many TSK FRBMs perform a weighted average to compute the output, additive FRBMs are also a common choice. They have been used in a large number of applications, for example [7,11,14,25].

When applied to model or forecast a univariate time series $\{y_t\}$, the rules of a TSK FRBM are expressed as

$$\begin{aligned}&\text{IF } y_{t-1} \text{ IS } A_1 \text{ AND } y_{t-2} \text{ IS } A_2 \text{ AND, } \ldots, \text{ AND } y_{t-p} \text{ IS } A_p\\ &\quad \text{THEN } y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_p y_{t-p}.\end{aligned} \tag{12}$$

In this rule, all the variables y_{t-i} are lagged values of the time series, $\{y_t\}$.

3.1. The AR model and the TSK fuzzy rules

Fuzzy rules are the core element of fuzzy systems. When applied to Time Series, as seen in Eq. (12), fuzzy rules can describe the relationship between the lagged variables in some parts of the state-space. From [1], a close look into this equation suggests that when used for time series modeling, a TSK fuzzy rule can be seen as a local AR model, applied on the state-space subset defined by the rule antecedent.

This connection between the two models opened the possibility of an exchange of knowledge from one field to another, enabling us to apply what we know about AR models to fuzzy rules and vice versa. From the point of view of the Box–Jenkins models, this kind of fuzzy rules represents a local AR model which is applied only when some conditions hold. These conditions are given by the terms in the rule antecedent, and are expressed as the fuzzy

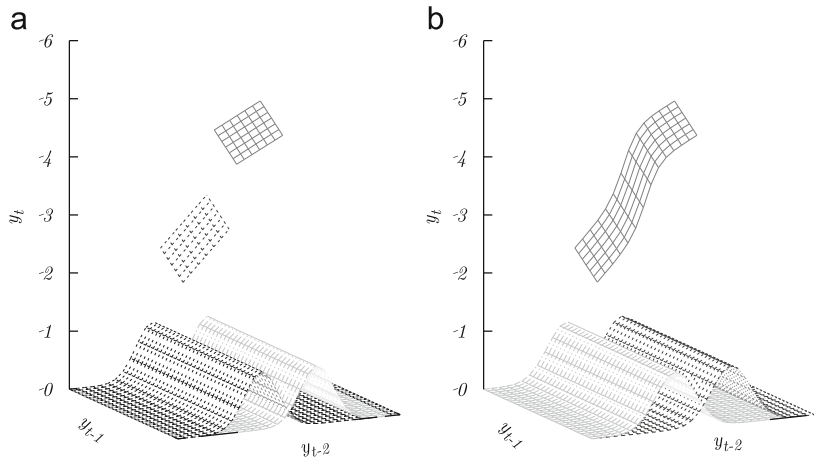

Fig. 3. (a) Two local AR models (or two fuzzy rules) (b) The STAR model (or the fuzzy inference system) derived from the two AR (or rules) shown in (a).

membership degree of the lagged variables to some fuzzy sets describing parts of the state-space domain. This scheme is closely related to the structure of the Threshold Autoregressive family of models, as shown below.

To stress the stochastic nature of fuzzy rules and their relationship with AR models, a random shock term ε_t should be included in the expression of the rule:

$$\begin{aligned} &\text{IF } y_{t-p} \text{ IS } A_1 \text{ AND, } \ldots, \text{ AND } y_{t-1} \text{ IS } A_p \\ &\quad \text{THEN } y_t^* = b_1 y_{t-p} + \ldots + b_p y_{t-1} + b_{p+1} + \varepsilon_t \end{aligned} \tag{13}$$

to obtain what might be called *fuzzy autoregressive rules*.

3.2. STAR model and fuzzy rule-based models

After the previous result, we went further in the exploration of the relationships between threshold models and fuzzy logic-based models. On the one hand, we have seen that AR models are good linear models applicable to prediction problems. As well, we know that a TAR model is basically a set of local AR models, and that it allows for some nonlinearity in its computations. On the other hand, we have seen above how a fuzzy rule relates to an AR model. Knowing that fuzzy rule-based models contain sets of fuzzy rules, we were interested in considering the relationship existing between threshold models and fuzzy rule-based models.

It is rather clear that there is some parallelism between the two aforementioned families of models. At a high level, models from both sides are composed of a set of elements (AR–fuzzy rules) which happen to be closely related, as stated above. On a lower level, both families of models rely on building a hyper-surface on the state-space which tries to model the relationship between the lagged variables of a time series. Moreover, both define this hyper-surface as the composition of hyper-planes which apply only in certain parts of the state-space.

This can be seen clearly in Fig. 3, which shows the graphical representation of the fuzzy inference system or the STAR model. On the left side, we can see the graphical representation of the skeleton of two linear models of type $y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \varepsilon_t$, together with the representation of two Gaussian membership/transition functions. On the right side, the two local linear models (or fuzzy rules) are linked through the smoothing functions into a single surface which represents a STAR model or a FRBM.

Indeed, in [1] it was proved that the STAR model is functionally equivalent to an additive TSK FRBM with only one term in the rule antecedents.

4. A linearity test for FRBMs

As stated before, a fundamental objection argued by scientists with a classical statistical background against soft computing models in general and neural networks and FRBM in particular was the lack of a sound theory behind them.

Not being able to prove *a priori* if such models had good statistical properties (related to their much praised "black-box" condition) prevented them to be accepted by wide parts of the scientific community despite its good performance in practical situations. Fuzzy-related researchers' and practitioners' attitude towards this has usually been to work from an engineering point of view and to further extend the practical applications of the models and methods in hope that their empirical benefits were at some point good enough as to finally convince the scientific community.

The results presented in Section 3 have an immediate impact on this question, as they permit the derivation of a statistical approach to a family of soft computing models, namely the FRBM family, considering them as nonlinear time series models.

This includes *a priori* proofs of their statistical properties, such as stationarity or identifiability, which will throw some light on their inner behavior. Also, the use of log-likelihood based estimation methods allow us to guarantee existence, convergence, consistence and asymptotic normality of the estimators. These properties are a must for a statistical model to be accepted. Finally, the development of linearity tests grant the ability to decide, based on the data, if a series can be modeled with a single linear autoregressive model or if an FRBM seems appropriate instead.

Before deriving these tests, a word on notation must be said. In the standard FRBM framework, the residuals are considered as an information source about the "goodness of fit" of the model. They are looked at once the model is built, as they are the basis for computing the so-called *error measures*: mean squared error, mean average error and so on.

In the statistical field, on the other hand, the time series formed by the residuals, $\{\varepsilon_t\}$, is a fundamental piece of the modeling process, and as such it is always included in the definition of the models. Hence, we will redefine the additive TSK FRBM, Eq. (11), in the time series framework as

$$y_t = G(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t = \sum_{i=1}^{r} \mathbf{b}_i' \mathbf{x}_t \cdot \omega_i(\mathbf{x}_t; \boldsymbol{\psi}_\omega) + \varepsilon_t, \tag{14}$$

where $\boldsymbol{\psi}$ is the parameter vector, including the consequent (linear) parameters, $\boldsymbol{\psi}_p = (\mathbf{b}_1, \ldots, \mathbf{b}_r)$ and the antecedent (nonlinear) parameters, $\boldsymbol{\psi}_\omega$, whose number depends on the type of membership function, μ_A, used. The residuals, ε_t, are henceforth included in the definition of the FRBM.

In this section, we will consider membership functions of Gaussian type, being the most common derivable membership functions used in this context. It is usually expressed as in Eq. (9) but we will rewrite it as

$$\mu(\mathbf{x}_t; \gamma, c) = \exp(-\gamma(x - c)^2). \tag{15}$$

Since FRBMs can be seen as nonlinear regression models, the standard procedures for testing parameter significance, like LM-tests, should be applicable, in principle. To perform these tests, however, the asymptotic distribution of the model parameters must be known. This issue is dealt with in Section 6, where it is shown that the parameters of an FRBM are asymptotically normal.

In the fuzzy literature, however, no attention has been paid to hypothesis testing up to now. While it is obvious that a linear time series should be modeled with a linear model, i.e. a single (default) rule, to our knowledge there is no testing procedure to avoid the mistake of using highly complex structures to model simple problems.

Next we propose a statistical test to decide if a problem can be solved using a linear model or if we need a combination of rules to model it. Let us suppose that we have an FRBM composed of a single linear model which applies to the whole input space:

$$y_t = \mathbf{b}_0' \mathbf{x}_t + \varepsilon_t. \tag{16}$$

Now we want to know if the use of an extra rule with Gaussian membership function would increase the performance of the model. We would add such rule as follows:

$$y_t = \mathbf{b}_0' \mathbf{x}_t + \mathbf{b}_1' \mathbf{x}_t \omega(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t. \tag{17}$$

If our FRBM uses Gaussian membership functions, with $\boldsymbol{\psi} = [\gamma, \mathbf{c}]$, we might rewrite it as

$$y_t = \mathbf{b}_0' \mathbf{x}_t + \mathbf{b}_1' \mathbf{x}_t \prod_{i=1}^{p} \exp(-\gamma(x_i - c_i)^2) + \varepsilon_t.$$

Our goal is to test for the significance of the extra rule, so in this case an appropriate null hypothesis could be

$$\mathrm{H}_0 : \gamma = 0, \tag{18}$$

being the alternative $\mathrm{H}_1 : \gamma > 0$. Hypothesis (18) opens up the possibility of studying linearity in the Lagrange multiplier (LM) testing framework. Under this null hypothesis, the contribution of the extra rule is identically equal to a constant and merges with the intercept b_{00} of the default rule, that is, the rule is not necessary.

We assume that, under (18), the maximum likelihood estimators of the parameters of (15) are asymptotically normal and hence can be estimated consistently (as granted by Theorem A.2, Appendix A).

As it was thoroughly discussed in [2], model (17) is only identifiable under the alternative hypothesis, i.e., if the null is true, the parameters are not locally unique and thus the estimator does not follow an asymptotic normal distribution. This issue is known as the problem of "hypothesis testing when a nuisance parameter is present only under the alternative", and was first studied by [9]. In this situation the test statistic of the LM-test does not follow a known distribution and thus the standard asymptotic distribution theory for the likelihood ratio is not available.

However, we can avoid this difficulty and obtain a χ^2-statistic by following the method first suggested in [15] and then widely applied to neural network-based models by [22,16,20] amongst others. This method proposes the expansion of the expression of the firing strength of a fuzzy rule into a Taylor series around the null hypothesis $\gamma = 0$:

$$\omega(\mathbf{x}_t; \gamma, \mathbf{c}) \approx \omega(\mathbf{x}_t; 0, \mathbf{c}) + \left.\frac{\partial \omega}{\partial \gamma}\right|_{\gamma=0} \gamma + R(\mathbf{x}_t; \gamma, \mathbf{c}) = \gamma \sum_{i=1}^{p} (x_i - c_i)^2 + R(\mathbf{x}_t; \gamma, \mathbf{c}), \tag{19}$$

which for the expression of the contribution of the extra rule yields

$$C \approx \mathbf{b}_1' \mathbf{x}_t \left[\gamma \sum (x_i - c_i)^2\right] = \sum_{i=1}^{p} \theta_i x_i + \sum_{i=1}^{p} \sum_{j=i}^{p} \theta_{ij} x_i x_j + \sum_{i=1}^{p} \sum_{j=i}^{p} \sum_{k=j}^{p} \theta_{ijk} x_i x_j x_k.$$

In this case, contrary to what happens when using the sigmoid membership function (as in the STAR model [15]), the first order Taylor approximation is enough for our needs, as all the θ_i, θ_{ij}, θ_{ijk} depend on the intercept, b_{10}, of (17). The first linear term merges with the system's default rule, while the remainder of the Taylor expansion adds up to the error term, becoming $\varepsilon^\star = \varepsilon + \mathbf{b}_1 \mathbf{x}_t R(\mathbf{x}_t; \gamma, \mathbf{c})$, which means that $\varepsilon^\star = \varepsilon$ under the null. Thus the expansion results in the following model:

$$y_t = \boldsymbol{\pi}' \mathbf{x}_t + \sum_{i=1}^{p} \sum_{j=1}^{p} \theta_{ij} x_i x_j + \sum_{i=1}^{p} \sum_{j=i}^{p} \sum_{k=j}^{p} \theta_{ijk} x_i x_j x_k + \varepsilon_t^\star. \tag{20}$$

The null hypothesis can hence be defined as

$$\mathrm{H}_0 : \theta_{ij} = 0 \wedge \theta_{ijk} = 0 \quad \forall i, j, k \in 1, \ldots, q. \tag{21}$$

This null hypothesis circumvents the identification problem, and allows us to obtain a statistical test concerning the use of the extra rule. This test is based on the local approximation to the log-likelihood for observation t, which takes the form (ς is the variance of ε)

$$l_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \varsigma^2 - \frac{1}{2\varsigma^2} \left\{ y_t - \boldsymbol{\pi}' \mathbf{x}_t - \sum_{i=1}^{p} \sum_{j=1}^{p} \theta_{ij} x_i x_j - \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \theta_{ijk} x_i x_j x_k \right\}^2. \tag{22}$$

Following [16], we must rely on the following assumptions:

Assumption 4.1. The $((r+1) \times 1)$ parameter vector defined by $[\boldsymbol{\psi}', \varsigma^2]'$ is an interior point of the compact parameter space $\boldsymbol{\Psi}$ which is a subspace of $\mathbb{R}^r \times \mathbb{R}^+$, the r dimensional Euclidean space.

Assumption 4.2. Under the null hypothesis, the data generating process (DGP) for the sequence of scalar real valued observations $\{y_t\}_{t=1}^T$ is an ergodic stochastic process, with true parameter vector $\boldsymbol{\psi} \in \boldsymbol{\Psi}$.

Assumption 4.3. $E|y_{t-i}|^{\delta} < \infty, \forall i \in \{1, \ldots, p\}$ for some $\delta > 8$.

Under H^0 and Assumptions 4.1–4.3 we can compute the standard Lagrange multiplier or score-type test statistic given by

$$\mathrm{LM} = \frac{1}{\hat{\sigma}^2} \sum_{t=1}^{T} \hat{\varepsilon}\hat{\boldsymbol{\tau}}_t' \times \left\{ \sum_{t=1}^{T} \hat{\boldsymbol{\tau}}_t \hat{\boldsymbol{\tau}}_t' - \sum_{t=1}^{T} \hat{\boldsymbol{\tau}}_t \hat{\mathbf{h}}_t' \times \left(\sum_{t=1}^{T} \hat{\mathbf{h}}_t' \hat{\mathbf{h}}_t \right)^{-1} \times \sum_{t=1}^{T} \hat{\mathbf{h}}_t \hat{\boldsymbol{\tau}}_t' \right\} \times \sum_{t=1}^{T} \hat{\boldsymbol{\tau}}_t' \hat{\varepsilon}, \tag{23}$$

where $\hat{\varepsilon} = y_t - \boldsymbol{\pi}' \mathbf{x}_t$ are the residuals estimated under the null hypothesis,

$$\hat{\mathbf{h}}_t = \left. \frac{\partial G(\mathbf{x}_t; \boldsymbol{\psi}_p, \boldsymbol{\psi}_\omega)}{\partial \hat{\boldsymbol{\psi}}_p \partial \hat{\boldsymbol{\psi}}_\omega} \right|_{\boldsymbol{\psi}_p = \hat{\boldsymbol{\psi}}_p \wedge \boldsymbol{\psi}_\omega = \hat{\boldsymbol{\psi}}_\omega} \tag{24}$$

is the gradient of the model and $\hat{\boldsymbol{\tau}}_t$ contains all the nonlinear regressors in (20), with $\|\hat{\boldsymbol{\tau}}_t\| = m$. This statistic has an asymptotic χ^2 distribution with m degrees of freedom.

Although it might seem complicated at first sight, this test can be easily carried out in stages:

1. Regress y_t on $\mathbf{x}_t$ and compute the residual sum of squares $SSR_0 = \sum_{t=1}^{T} \hat{\varsigma}_t^{\,2}$.
2. Regress $\hat{\varsigma}_t$ on $\mathbf{x}_t$ and on the m nonlinear regressors of (20). Compute the residual sum of squares $SSR_1 = \sum_{t=1}^{T} \hat{\tau}_t^{\,2}$.
3. Compute the χ^2 statistic

$$\mathrm{LM}_{\chi^2}^{l} = T \frac{SSR_0 - SSR_1}{SSR_0}$$

or the F version of the test

$$\mathrm{LM}_{F}^{l} = \frac{(SSR_0 - SSR_1)}{m} \left(\frac{SSR_1}{(T - p - 1 - m)} \right)^{-1}.$$

If the value of the test statistic exceeds the appropriate value of the χ^2 or F distribution, the null hypothesis is rejected.

There is a publicly available implementation of this test in the statistical language `R`, www.r-project.org, where the linearity test has been implemented as part of the package `tsDyn` [17]. Having loaded the package, applying the test is as simple as calling the function `linearityTest` with the series under study as the main argument.

5. Power analysis

The use of synthetic datasets has been recently studied in the framework of Soft Computing. For a detailed state-of-the-art, see [5]. Nonetheless, in the statistical field, it is a common practice to use this type of experiments to check the modeling capabilities of the proposals.

The basic assumption is that any series is considered to be generated by a usually unknown data generating process (DGP) to which a noise component is added:

$$y_t = G(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t. \tag{25}$$

As a reverse result of this, to generate an artificial time series, we need to define a DGP and a noise distribution, whose sum in iterative application will produce the data. This artificial series could then be studied under the chosen modeling scheme, identifying and estimating a model for it. If the parameters of this model are (or tend to be) equal to the parameters of the original DGP, we obtain a clear evidence that the modeling scheme is correct.

In order to simulate a series according to the aforementioned basic assumption, we must go back again to the expression of the general model, Eq. (25). The first part of the right hand side of that expression is called in this context the *model skeleton*, and of course is the part which is to be modeled. Having defined a model skeleton or DGP, we generate the series by seeding a random starting point $\mathbf{x}_{t_0}$ and successively obtaining the $y_t, t = 1, \ldots, T$, by applying

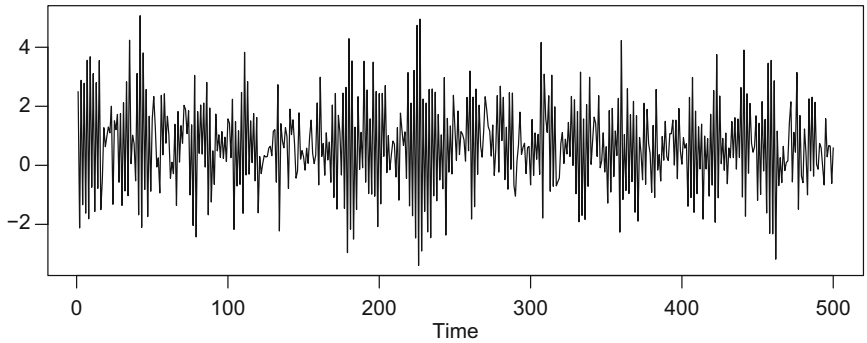

Fig. 4. Generated series for model (26), stationary linear autoregressive model.

the skeleton function and adding an n.i.d. value given by the random series ε_t. It is usually a good idea to discard the first N observations to avoid initialization effects.

In this study, we generated 10 synthetic time series. Some of them are similar to the ones used by [16], that we reuse in order to test them in the FRBM framework.

We start by simulating five stationary linear autoregressive models:

$$y_t = 0.8 - 0.5y_{t-1} + 0.3y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1^2), \tag{26}$$

$$y_t = -0.1 + 0.2y_{t-1} + 0.2y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.5^2), \tag{27}$$

$$y_t = -0.4 + 0.7y_{t-1} + 0.1y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.2^2), \tag{28}$$

$$y_t = 0.3 - 0.4y_{t-1} - 0.5y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1.9^2), \tag{29}$$

$$y_t = 0.5 + 0.2y_{t-1} + 0.6y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.9^2), \tag{30}$$

whose simple formulation produces a series as the one shown in Fig. 4 (corresponding to model (26)).

Knowing that these series are linear, we first wanted to check if the null hypothesis of the linearity test would be accepted or not. By using the skeleton and the random noise series, we simulated 500 replications of the model (containing 500 observations each) and applied the test to them. [2]

For a significance level of 0.05, the results are shown in the first five rows of Table 1. The first and second columns show the number of series for which the null hypothesis was accepted and rejected, respectively. The third column shows the percentage of correct answers and the fourth shows the average p-value obtained. As we can see, in the five examples the test properly accepted the null hypothesis in more than 90% of the cases.

Once we know empirically that the test is robust against Type I errors, we turn to investigate Type II errors, i.e., the null hypothesis of linearity not being rejected when dealing with nonlinear series. In order to do so, we simulated the following five models:

$$y_t = 1.8y_{t-1} - 1.06y_{t-2} + (0.02 - 0.9y_{t-1} + 0.795y_{t-2}) \times \mu(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.02^2) \tag{31}$$

with $\boldsymbol{\psi} = [\gamma, \mathbf{c}] = [3, (1.2, 0.7)]$:

$$\begin{aligned} y_t = {} & -0.1 + 0.3y_{t-1} + 0.2y_{t-2} + (-1.2y_{t-1} + 0.5y_{t-2}) \times \mu_1(\mathbf{x}_t; \boldsymbol{\psi}_1) \\ & + (1.8y_{t-1} - 1.2y_{t-2}) \times \mu_2(\mathbf{x}_t; \boldsymbol{\psi}_2) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.5^2) \end{aligned} \tag{32}$$

with $\boldsymbol{\psi}_1 = [\gamma_1, \mathbf{c}_1] = [2, (2.0, 0.1)]$ and $\boldsymbol{\psi}_2 = [\gamma_2, \mathbf{c}_2] = [4.3, (0.31, 1.5)]$:

$$y_t = 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (-0.5 - 1.2y_{t-1} + 0.8y_{t-2}) \times \mu(\mathbf{x}_t; \boldsymbol{\psi}) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.5^2) \tag{33}$$

[2] All the experiments were carried out in the statistical programming language `R` using the package `tsDyn`.

Table 1
Results of the linearity tests applied to the synthetic series.

Model	Accept H_0	Reject H_0	% correct	Average p-value
Eq. (26)	461	39	92.2	0.5490
Eq. (27)	476	24	95.2	0.5047
Eq. (28)	467	33	93.4	0.5254
Eq. (29)	471	29	94.2	0.4825
Eq. (30)	481	19	96.2	0.5283
Eq. (31)	6	494	98.8	0.0029
Eq. (32)	2	498	99.6	0.0009
Eq. (33)	0	500	100	0.00001
Eq. (34)	0	500	100	1.18e−29
Eq. (35)	0	500	100	2.75e−8

with $\boldsymbol{\psi} = [\gamma, \mathbf{c}] = [11.31, (0.7071, -0.7071)]$:

$$y_t = 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (1.5 - 0.6y_{t-1} - 0.3y_{t-2}) \times \mu_1(\mathbf{x}_t; \boldsymbol{\psi}_1) + (-0.5 - 1.2y_{t-1} + 0.7y_{t-2}) \times \mu_2(\mathbf{x}_t; \boldsymbol{\psi}_2) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 1^2) \tag{34}$$

where $\boldsymbol{\psi}_1 = [\gamma_1, \mathbf{c}_1] = [8.49, (0.7071, -0.7071)]$ and $\boldsymbol{\psi}_2 = [\gamma_2, \mathbf{c}_2] = [8.49, (-0.7071, 0.7071)]$:

$$\begin{aligned} y_t = {} & 0.5 + 0.8y_{t-1} - 0.2y_{t-2} + (1.5 - 0.6y_{t-1} - 0.3y_{t-2}) \times \mu_1(\mathbf{x}_t; \boldsymbol{\psi}_1) \\ & + (0.2 + 0.3y_{t-1} - 0.9y_{t-2}) \times \mu_2(\mathbf{x}_t; \boldsymbol{\psi}_2) + (-1.2 + 0.6y_{t-1} + 0.8y_{t-2}) \times \mu_3(\mathbf{x}_t; \boldsymbol{\psi}_3) \\ & + (-0.5 - 1.2y_{t-1} + 0.7y_{t-2}) \times \mu_4(\mathbf{x}_t; \boldsymbol{\psi}_4) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, 0.2^2), \end{aligned} \tag{35}$$

where

$$\begin{aligned} \boldsymbol{\psi}_1 &= [\gamma_1, \mathbf{c}_1] = [2.34, (-0.4, -0.7)], \\ \boldsymbol{\psi}_2 &= [\gamma_2, \mathbf{c}_2] = [2.34, (-0.7, 0.4)], \\ \boldsymbol{\psi}_3 &= [\gamma_3, \mathbf{c}_3] = [4.23, (0.11, -0.65)], \\ \boldsymbol{\psi}_4 &= [\gamma_4, \mathbf{c}_4] = [4.23, (0.65, -0.11)]. \end{aligned}$$

Again, for each of these models, we simulated 500 series, of 500 observations each, and applied the test to each of them. For a significance level of 0.05, the five last rows of Table 1 show that the test is very good against Type II errors, as the null hypothesis was not rejected in only eight series over a total of 2500.

In order to further assess the goodness of the test, we studied its power with respect to the sample size. For a nominal size (significance) of 0.05, the theoretical power of the test is shown in Fig. 5. The classification proposed by [8] uses the *effect size* as a way to express the strength of the relationship between two variables. According to this classification and the curves shown in Fig. 5, if we have a "small" effect size (0.2), the test shows good power already when we use a sample size of 300. For a "medium" effect size (0.5), the test is powerful already with 50 samples, and with a "large" effect size (0.8), even less than 25 samples are enough to obtain good test power.

Fig. 6 shows empirical power with respect to nominal size or significance for model (31). This model was chosen because, amongst the nonlinear models, it is less complex and hence it is easier to mistakenly accept the null hypothesis. This is also seen in Table 1, where model (31) gets the worse percentage of correct rejections amongst the nonlinear models. However, the empirical power shown in the figure is clearly very strong (over 0.99 for the standard nominal size 0.05), so it is clear that the test is good enough even for not very complex models.

To conclude, in order to further illustrate the usefulness of the test, we will show how neglecting the information provided by it might lead to a waste of resources and time.

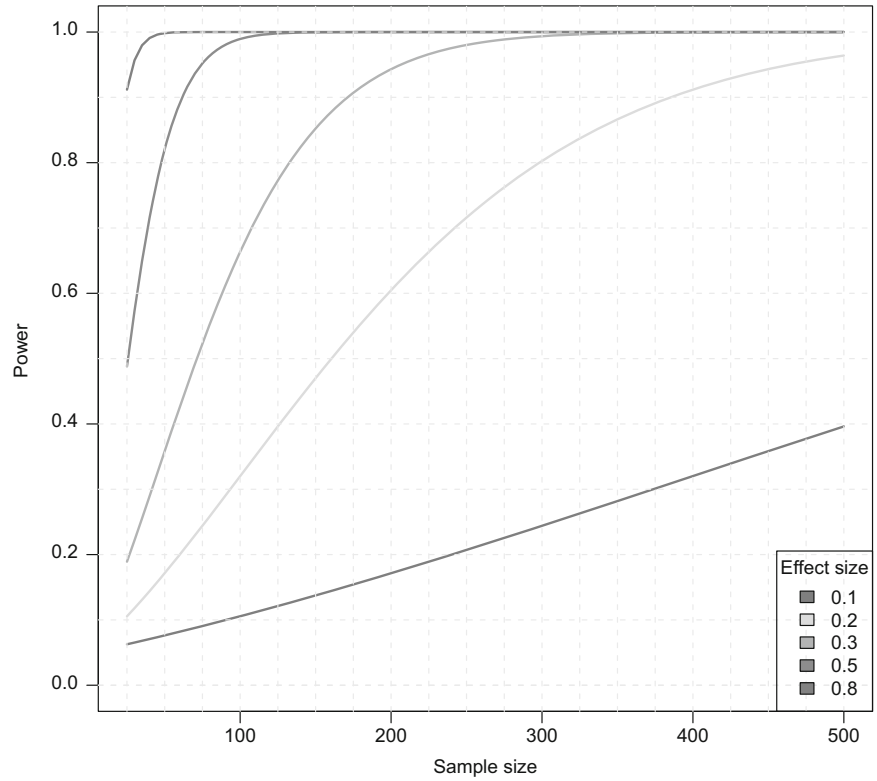

Fig. 5. Power estimation for linearity test.

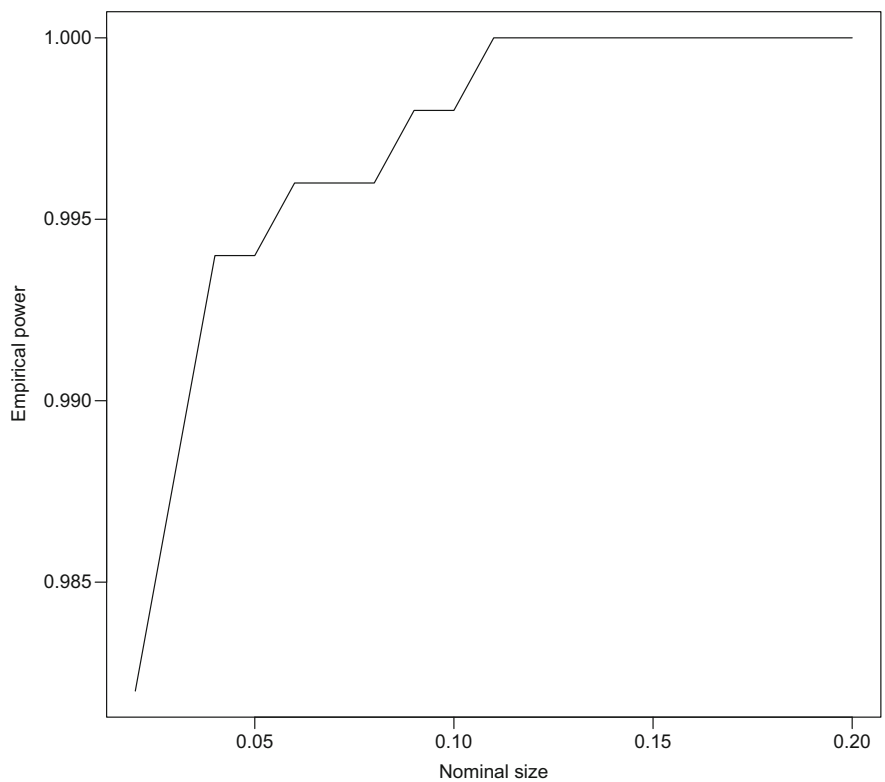

Fig. 6. Empirical power versus nominal size for the linearity test applied on model (31).

Let us assume that we want to model the data shown in Fig. 4 by using two lagged variables, y_{t-1} and y_{t-2}. At first sight this series looks quite complex, and we might want to try to model it using a fuzzy rule-based model. If we choose to apply a Wang and Mendel [26] grid partition strategy to define the structure of the model, and we use three Gaussian-based labels for each fuzzy variable, we will end up with a model having nine rules and a total of $3 \times 9 + 2 \times 6 = 39$ parameters.

However, Wang and Mendel algorithm is known for not being very parsimonious and we might use a better algorithm such as substractive clustering to define the structure of the model. In that case, using the default values for the algorithm's parameters, we get a 3 rule model with a total of $3 \times 3 + 2 \times 6 = 21$ parameters. After some tuning of the substractive clustering algorithm, the smallest model we get is a 2 rule model with $3 \times 2 + 2 \times 4 = 14$ parameters.

After accepting the null hypothesis of linearity, we know that the series shown in Fig. 4 is generated (and hence best modeled) by a linear model with just three parameters. A linear regression is sufficient to obtain a good model (26).

6. Conclusions

As a new step towards a statistical framework for fuzzy rule-based modeling of time series, this paper introduces a test for linearity and some results concerning the maximum likelihood estimator. These contributions are relevant both in a theoretical and practical point of view.

On the one hand, in deriving the theoretical properties of the estimator and the test itself, we move closer to a convincing statistical foundation for FRBM which will surely lead to a better understanding of the model and a deeper insight into its behavior. As a side-effect, these contributions could also be another argument to temper the traditional skepticism amongst Statistics researchers concerning FRBM.

On the other hand, the practical implications of the contributions shown above are clear. The proposed linearity test, of a Lagrange multiplier type, allows for an empirical determination of the nonlinearity of a series. This is an important issue which should be considered by any researcher before trying to model a time series using a FRBM. It is clear that it would be a loss of time and resources to model a linear time series with a highly nonlinear model as the FRBM, and hence it is something to be avoided.

Finally, this work is part of a continued effort to bring statistical results, methods and tools, to the fuzzy rule-based time series modeling. Further developments are expected to come, aiming at a complete statistical approach to time series modeling and forecasting through FRBM.

Acknowledgments

The research included in this work is part of the first author's PhD dissertation at the Department of Computer Science and Artificial Intelligence of the University of Granada. It has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government through Research Grants Ref. TIN2009-14575 and CIT-460000-2009-46.

Appendix A. Estimation procedures. Properties of the estimator

The linearity test developed above requires the estimation of the FRBM. In this section we will focus on this problem.

There is a growing number of algorithms in the literature for estimating the parameters of FRBM and Neural Network based models. Following the reasons argued in [22,16], we choose to estimate the parameters of the model by maximum likelihood, making use of the assumptions made previously on ε_t.

Estimation through maximum likelihood has been applied to Neural Networks but it is not so common to find it applied to FRBM. This is another idea that we borrow from classical statistic time series analysis, and it makes it feasible to obtain an idea of the uncertainty in the parameter estimates through asymptotic standard deviation estimates, which is something hardly possible through metaheuristic algorithms. It may be argued, though, that maximum likelihood estimation of neural network models is most likely to lead to convergence problems, and that penalizing the log-likelihood function is a necessary precondition for obtaining satisfactory results. Notwithstanding, in this case, unidentifiable models are not estimated (a main reason for penalizing the log-likelihood), and furthermore the initial values for the parameters are carefully chosen.

If, as we assume here, ε_t is a Gaussian white noise with zero mean and finite variance, $\varepsilon_t \backsim \text{NID}(0, \sigma^2)$, the maximum likelihood is equivalent to nonlinear least squares. Hence, the parameter vector $\boldsymbol{\psi}$ of the model is

estimated as

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}} Q_T(\boldsymbol{\psi}) = \arg\min_{\boldsymbol{\psi}} \sum_{t=1}^{T} (y_t - G(\mathbf{x}_t; \boldsymbol{\psi}))^2. \tag{36}$$

The least squares estimator (LSE) defined by (36) belongs to the class of M estimators considered by [19]. We next discuss the conditions that guarantee existence, consistency and asymptotic normality of the LSE.

A.1. Existence of the estimator

Following [20], the existence of the LSE estimator is based in Lemma 2 of [10], which establishes the existence under certain conditions of continuity and measurability on the mean squared error (MSE) function.

Theorem A.1. *The additive TSK FRBM satisfies the following conditions and the LSE exists*:

1. *For each* $\mathbf{x}_t \in \mathbf{X}$, *function* $G(\mathbf{x}_t; \boldsymbol{\psi})$ *is continuous in a compact subset* $\boldsymbol{\Psi}$ *of the Euclidean space*.
2. *For each* $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, *function* $G(\mathbf{x}_t; \boldsymbol{\psi})$ *is measurable in space* $\mathbf{X}$.
3. ε_t *are independent and identically distributed errors with mean* 0 *and variance* σ^2.

Proof. Lemma 2 of [10] shows that conditions 1–3 in Theorem A.1 are sufficient to guarantee the existence (and measurability) of the LSE. To apply this result to the FRBM, we need to check whether these conditions are satisfied by the model.

Condition 3 of Theorem A.1 was already assumed when defining the model. It is easy to prove in our case that $G(\mathbf{x}_t; \boldsymbol{\psi})$ is continuous in the parameter vector $\boldsymbol{\psi}$. This follows from the fact that $p(\mathbf{x}_t; \boldsymbol{\psi}_\mu)$ and $\mu(\mathbf{x}_t; \boldsymbol{\psi}_p)$ depend continuously on $\boldsymbol{\psi}_p$ and $\boldsymbol{\psi}_\mu$ for each value of $\mathbf{x}_t$. Similarly we can see that $G(\mathbf{x}_t; \boldsymbol{\psi})$ is continuous in $\mathbf{x}_t$ and thus is measurable, for each fixed value of the parameter vector $\boldsymbol{\psi}$. Thus, conditions 1 and 2 are also satisfied. □

A.2. Consistence and asymptotic normality of the estimator

White [28,29] established the conditions that guarantee strong consistency of the LSE. In the context of stationary time series models, the conditions that ensure (almost certain) consistency have been established in [27,30]. Now, as [2] guarantees the global identifiability of the model, we can prove existence, consistency and asymptotic normality of the FRBM estimators.

Theorem A.2. *Under Assumptions* 4.1, 4.2 *and Theorem* 1 *of* [2], *the maximum likelihood estimator* $\hat{\boldsymbol{\psi}}$ *is almost surely consistent for* $\boldsymbol{\psi}$ *and*

$$\sqrt{T}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{D} N\left(0, -\operatorname*{plim}_{T\to\infty} \mathbf{A}(\boldsymbol{\psi}^{-1})\right), \tag{.37}$$

where

$$\mathbf{A}(\boldsymbol{\psi}^{-1}) = \left(\frac{1}{\sigma^2 T}\right)\left(\frac{\partial^2 Q_T(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right).$$

Proof. To prove consistency, we will make use of Theorem 3.5 of [27] and follow [16]. We must show that the assumptions of that theorem are fulfilled in the case of the FRBM:

Assumptions 2.1 and 2.3, related to the probability space and to the density functions, are trivial. Let $q(\mathbf{x}_t; \boldsymbol{\psi}) = (y_t - G(x_t; \boldsymbol{\psi}))^2$. Assumption 3.1a states that for each $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, $-\mathrm{E}(q(\mathbf{x}_t; \boldsymbol{\psi}))$ exists and is finite for $t = 1, \ldots, T$. Under Assumption 4.2 and the fact that ε_t is a zero mean normally distributed random variable with finite variance, hence, k-integrable, Assumption 3.1a in [27] follows.

Assumption 3.1b states that $-\mathrm{E}(q(\mathbf{x}_t; \boldsymbol{\psi}))$ is continuous in $\boldsymbol{\Psi}$, $t = 1, \ldots, T$. Let $\boldsymbol{\psi} \to \boldsymbol{\psi}^*$, since for any t, $G(x_t; \boldsymbol{\psi})$ is continuous on $\boldsymbol{\Psi}$, then $q(\mathbf{x}_t; \boldsymbol{\psi}) \to q(\mathbf{x}_t; \boldsymbol{\psi}^*)$, $\forall t$ (point-wise convergence). From the continuity of $G(x_t; \boldsymbol{\psi})$ on the compact set $\boldsymbol{\Psi}$, we have uniform continuity and we obtain that $q(\mathbf{x}_t; \boldsymbol{\psi})$ is dominated by an integrable function dF.

Then, by Lebesgue's dominated convergence theorem, we get $\int q(\mathbf{x}_t;\boldsymbol{\psi})\,dF \to \int q(\mathbf{x}_t;\boldsymbol{\psi}^*)\,dF$ and $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ is continuous.

Assumption 3.1c states that $-\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ obeys the strong (weak) law of large numbers (ULLN). Lemma 2 of [19] guarantees that $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ obeys the strong law of large numbers. The set of hypothesis (b) of this lemma is satisfied:

1. we deal with an ergodic process,
2. from the continuity of $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}))$ and from the compactness of $\boldsymbol{\Psi}$ we have that $\inf \mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi})) = \mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}^*))$ for $\boldsymbol{\psi}^* \in \boldsymbol{\Psi}$, and with Assumption 3.1a in [27] we may guarantee that $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}^*))$ exists and is finite, getting that $\mathrm{E}(q(\mathbf{x}_t;\boldsymbol{\psi}^)) > -\infty$.

Assumption 3.2 is related to the unique identifiability of $\boldsymbol{\psi}^*$, which is guaranteed by [2].

Now, to prove normality, we will use Theorem 6.4 of [27] and will check its assumptions. Assumptions 2.1, 2.3 and 3.1 follow from the proof of consistency showed above. Assumptions 3.2 and 3.6 follow from the fact that $G(x_t;\boldsymbol{\psi})$ is continuously differentiable of order 2 on $\boldsymbol{\psi}$ in the compact space $\boldsymbol{\Psi}$.

In order to check Assumptions 3.7a and 3.8a we have to prove that $\mathrm{E}(\nabla Q_T(\boldsymbol{\psi})) < \infty$ and $\mathrm{E}(\nabla^2 Q_T(\boldsymbol{\psi})) < \infty$, $\forall T$. The expected gradient and the expected Hessian of $Q_T(\boldsymbol{\psi})$ are given by

$$\mathrm{E}(\nabla Q_T(\boldsymbol{\psi})) = 2\mathrm{E}(\nabla G(x_t;\boldsymbol{\psi})(y_t - G(x_t;\boldsymbol{\psi})))$$

and

$$\mathrm{E}(\nabla^2 Q_T(\boldsymbol{\psi})) = 2\mathrm{E}(\nabla G(x_t;\boldsymbol{\psi})\nabla G(x_t;\boldsymbol{\psi})' - \nabla^2 G(x_t;\boldsymbol{\psi})(y_t - G(x_t;\boldsymbol{\psi})))$$

respectively. Assumptions 3.7a and 3.8a follow considering the normality condition on ε_t, the properties of the function $G(x_t;\boldsymbol{\psi})$ and the fact that $\nabla G(x_t;\boldsymbol{\psi})$ and $\nabla^2 G(x_t;\boldsymbol{\psi})$ contain at most second order terms of $\mathbf{x}_t$.

Assumption 3.8c is guaranteed by the proof of consistency and the ULLN from [19].

Assumption 3.9 follows from the identifiability of the FRBM and the properties of function $G(x_t;\boldsymbol{\psi})$.

Assumption 6.1 requires using Theorem 2.4 of [29], by which we can show that $2\xi'\nabla G(x_t;\boldsymbol{\psi}^*)\varepsilon_t$ obeys the central limit theorem for some $(r \times 1)$ vector ξ, such that $\xi\xi' = 1$. Assumptions A(i) and A(iii) both hold because ε_t is Gaussian. Assumption A(ii) also holds with $V = 4\sigma^2\xi'\mathrm{E}(\nabla G(x_t;\boldsymbol{\psi}^*)\nabla' G(x_t;\boldsymbol{\psi}^*))$. Furthermore, since any measurable transformation of mixing processes is itself mixing (see [29, Lemma 2.1]), hence we have that $2\xi'\nabla G(x_t;\boldsymbol{\psi}^*)\varepsilon_t$ is a strong mixing sequence and obeys the central limit theorem. $\nabla Q_T(\boldsymbol{\psi})$ also obeys the CLT with covariance matrix $B_T^* = 4\sigma^2\mathrm{E}(\nabla G(x_t;\boldsymbol{\psi}^*)\nabla' G(x_t;\boldsymbol{\psi}^*)) = 2\sigma^2 A_T^*$, which is $O(1)$ and non-singular.

References

[1] M.J.L. Aznarte, J.M. Benítez, J.L. Castro, Smooth transition autoregressive models and fuzzy rule-based systems: functional equivalence and consequences, Fuzzy Sets and Systems 158 (24) (2007) 2734–2745.
[2] M.J.L. Aznarte, J.M.B. Sánchez, On the identifiability of TSK additive fuzzy rule-based models, in: Soft Methods for Integrated Uncertainty Modelling, Advances in Soft Computing, Vol. 6, Springer, Berlin, 2006, pp. 79–86.
[4] L. Bachelier, Calcul des Probabilités. Garthier-Villars, Paris, 1912.
[5] M. Basu, T.K. Ho (Eds.), Data Complexity in Pattern Recognition, Springer, Berlin, 2006.
[6] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, 1970.
[7] H.S. Byun, K.H. Lee, A decision support system for the selection of a rapid prototyping process using the modified TOPSIS method, International Journal of Advanced Manufacturing Technology 26 (11–12) (2005) 1338–1347.
[8] J. Cohen, Statistical Power analysis for the Behavioral Sciences, second ed., Lawrence Earlbaum Associates, London, 1988.
[9] R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternatives, Biometrika 74 (1) (1987) 33–43.
[10] R.I. Jennrich, Asymptotic properties of non-linear least squares estimators, The Annal of Mathematical Statistics 40 (1969) 633–643.
[11] R.I. John, P.R. Innocent, Modeling uncertainty in clinical diagnosis using fuzzy logic, IEEE Transactions on Systems, Man, and Cybernetics, Part B 35 (6) (2005) 1340–1350.
[12] A.Y. Khinchin, Korrelationstheorie der stationaren stochastischen Prozesse, Mathematical Annals 109 (604) (1934).
[13] A.N. Kolmogorov, Uber die Analytischen Methoden in der Wahrscheinlich-keitsrechnang, Mathematical Annals 104 (604) (1931).
[14] I. Lee, B. Kosko, W.F. Anderson, Modeling gunshot bruises in soft body armor with an adaptive fuzzy system, IEEE Transactions on Systems, Man, and Cybernetics, Part B 35 (6) (2005) 1374–1390.
[15] R. Luukkonen, P. Saikkonen, T. Teräsvirta, Testing linearity against smooth transition autoregressive models, Biometrika 75 (1988) 491–499.
[16] M.C. Medeiros, A. Veiga, A flexible coefficient smooth transition time series model, IEEE Transactions on Neural Networks 16 (1) (2005) 97–113.
[17] A.F. Di Narzo, J.L. Aznarte, M. Stigler, tsDyn: Time Series Analysis based on Dynamical Systems Theory, R package version 3.1-89, 2008.

[18] W. Pedrycz, F. Gomide, An Introduction to Fuzzy Sets: Analysis & Design, The MIT Press, Cambridge, MA, 1998.
[19] B.M. Potscher, I.R. Prucha, A class of partially adaptive one-step m-estimators for the non-linear regression model with dependent observations, Journal of Econometrics 32 (2) (1986) 219–251.
[20] M. Suarez-Farinas, C.E. Pedreira, M.C. Medeiros, Local global neural networks: a new approach for nonlinear time series modeling, Journal of the American Statistical Association 99 (2004) 1092–1107.
[21] T. Teräsvirta, Specification, estimation and evaluation of smooth transition autoregressive models, Journal of the American Statistical Association 89 (1994) 208–218.
[22] T. Teräsvirta, M.C. Medeiros, G. Rech, Building neural network models for time series: a statistical approach, Journal of Forecasting 25 (1) (2006) 49–75.
[23] H. Tong, On a threshold model, in: Pattern Recognition and Signal Processing, 1978.
[24] D. van Dijk, T. Teräsvirta, P. Hans Franses, Smooth transition autoregressive models—a survey of recent developments, Econometric Reviews 21 (1) (2002) 1–47.
[25] C. Fernando Vieira, L. Brito Palma, R. Neves da Silva, Robust fault diagnosis approach using analytical and knowledge based techniques applied to a water tank system, International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications 13 (4) (2005) 237–244.
[26] L. Wang, J.M. Mendel, Generating fuzzy rules by learning from examples, IEEE Transactions on Systems, Man and Cybernetics 22 (6) (1992) 1414–1427.
[27] H. White, Estimation, Inference and Specification Analysis, Cambridge University Press, Cambridge, 1994.
[28] H. White, Consequences and detection of misspecified nonlinear regression models, Journal of the American Statistical Association 76 (374) (1981) 419–433.
[29] H. White, I. Domowitz, Nonlinear regression with dependent observations, Econometrica 52 (1) (1984) 143–161.
[30] J.M. Wooldridge, Estimation and inference for dependent processes, in: R.F. Engle, D. McFadden (Eds.), Handbook of Econometrics, Vol. 4, Elsevier, Amsterdam, 1986, pp. 2639–2738, (Chapter 45).
[31] G.U. Yule, On the method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, Philosophical Transactions A 226 (1927) 267–298.