123

A Perturbative Approach to Novelty Detection in Autoregressive Models

Maurizio Filippone and Guido Sanguinetti

Abstract

We propose a new method to perform novelty detection in dynamical systems governed by linear autoregressive models. The method is based on a perturbative expansion to a statistical test whose leading term is the classical F-test, and whose $O(\frac{1}{n})$ correction can be approximated as a function of the number of training points and the model order alone. The method can be justified as an approximation to an information theoretic test. We demonstrate on several synthetic examples that the first correction to the F-test can dramatically improve the control over the false positive rate of the system. We also test the approach on some real time series data, demonstrating that the method still retains a good accuracy in detecting novelties.

Index Terms

novelty detection, autoregressive modeling, time series, statistical testing.

I. Introduction

Novelty detection is the problem of identifying unexpected/abnormal events in data sets based solely on normal examples. Due to its practical importance, the problem has drawn much attention and many approaches have been proposed, including neural networks [1], [2], extreme value statistic [3], information theory [4], kernel and support vector methods [5], [6], [7], frequentist [8] and Bayesian [9] non-parametric approaches (for a good review of statistical approaches for novelty detection see *e.g.* [10], [11]).

M. Filippone is with the Department of Computing Science of the University of Glasgow, Sir Alwyn Williams Building, Lilybank Gardens, G12 8QQ, Glasgow, UK. email: maurizio@dcs.gla.ac.uk.

G. Sanguinetti is with the School of Informatics of the University of Edinburgh, Informatics Forum, 10 Crichton Street, EH8 9AB, Edinburgh, UK. email: gsanguin@inf.ed.ac.uk

Most approaches rely on estimating some characteristics of the data distribution for the normal class from training data, and then use this distribution to define a measure of how novel a test point is. Due to the absence of information on the distribution of novel events, any novelty detection system will necessarily label some normal data as novel (false alarms), and an important characteristic of the system is its ability to accurately predict the rate with which false alarms will be raised. Depending on the application, it is important to balance the cost of letting some novelties be undetected, and the cost of raising too many false alarms.

Of particular interest is the problem of identifying novelties in time series due to its many applications ranging from condition monitoring in health-care [12], [13] to fault detection in engineering [14], [15], [16], [17]. We can distinguish between two subtly different goals when dealing with novelties. One is identifying novelties in order to mitigate their effect on parameter estimation. In other words, the outliers are assumed to contaminate the series under study and the goal is to cope with that in the modeling stage. In this kind of approach, the learning system can be set up off-line, and is often referred to outlier detection. In the context of time series, many approaches have been proposed with this aim [18], [19], [20], [21], [22], [23]. Another goal, instead, is learning a model from a set of data that is considered normal. In this case, the assumption is that the data used to train the learning system constitute the basis to build a model of normality and the decision process on test data is usually online and based on the model of normality [24], [25], [26], [27], [28]. Equally important is the distinction between event-based and model-based novelties. Event-based novelties, also known as *Additive Outliers* (AO), are single observations that deviate from the norm. Model-based novelties, also known as *Innovation Outliers* (IO), instead, arise when the system changes its behavior over time. Typically, when a model is constructed, this problem is translated in the identification of changes in the model parameters.

In this paper, we consider the online identification of event-based novelties in stationary linear autoregressive models with Gaussian noise. These constitute an important and broadly used class of dynamical systems where each observation is modeled as a linear combination of previous observations plus a normally distributed noise term. We approach this problem by using a perturbative approximation to an information theoretic measure, recently introduced in [4] for i.i.d. data. Specifically, we consider an approximation to the *information content* (for the definition see Section II) of a new element of the series. This is defined by considering the Kullback-Leibler (KL) divergence between the estimates of the distributions of the stochastic term obtained before and after the new point is considered.

The KL divergence has been used with success in many areas of statistics [29], from model identification to approximation in Bayesian inference. In this paper, we intend to use it to motivate our proposed measure

of the information content of a new data point.

We approximate such a measure by expanding it in powers of the inverse of the sample size about the true (unknown) parameter values. This procedure yields a modified F-test which is able to more accurately incorporate the variability introduced by finite sample size effects. We test the model on a variety of synthetic examples exploring a wide range of parameter values and model orders, and comparing with a number of competing methods. This indeed confirms that the proposed approach does lead to a tight control over the false alarm rate, while retaining a competitive ability to highlight true positives. We also explore the robustness of the method to violations of its assumptions, testing on data generated by non-Gaussian non-linear autoregressive models. The model is still shown to perform well at controlling the false positive rates, although the non-Gaussian nature of the data leads to a slightly over-conservative bias. We then test our model on two historic time series from an environmental and a financial data sets and show that our approach is still able to capture exceptional events corresponding to true event-based novelties.

In the i.i.d. case presented in [4], the information content of a new data point was measured using the KL divergence between the estimated distribution on the training data and the one when a new test point was added. The motivation for using the KL divergence to test for novelties lies in its connections with Neyman-Pearson lemma [30]. It has been shown that in the Gaussian case, such an approach is analytically tractable and yields a test that is related to the F-test [4]. In the case of mixture models, it has been necessary to resort to some approximations and sampling to obtain a test for novelties. In this paper, instead, we approximate the information content of a new data point by obtaining an analytical correction to an F-test on the stochastic terms. The motivation for applying the test on the stochastic terms, rather than the full process, is that it requires the computation of the KL divergence between univariate Gaussians, which is more easily evaluated than the multidimensional divergences that would be obtained using the full process.

The paper is organized as follows: in Section II we sketch the derivation of the proposed statistical test for novelty detection for linear autoregressive models; in Section III we show some experiments on synthetic and real data sets; in Section IV we draw the conclusions. The full derivation of the method is reported in the Appendix.

II. Statistical Testing for AR(d) time series with i.i.d. Gaussian noise

A. Parameter Estimation

Let us consider a time series $X = \{x_1, x_2, \ldots, x_n\}$. A *linear autoregressive model* of order d (AR(d)) describes each observation as a linear combination of d past observations plus a stochastic term. In other words, an AR(d) model can be written as:

$$x_{t+1} = \sum_{j=1}^{d} \alpha_j x_{t+1-j} + \varepsilon_{t+1} + \mu = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{x}_t + \varepsilon_{t+1} + \mu \tag{1}$$

having introduced the vectors:

$$\mathbf{x}_t = (x_t, x_{t-1}, \ldots, x_{t-d+1})$$

The d coefficients of the linear combination are contained in the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$. The terms ε_{t+1} are i.i.d. and distributed as a $\mathcal{N}(0, \gamma^2)$. The value μ allows to model series with non-zero mean. In the following, we will assume that the process is stable.

By imposing the first order stationarity of $\mathrm{E}[x_t]$, that is $m = \mathrm{E}[x_t]\ \forall t$:

$$\mathrm{E}[x_{t+1}] = \sum_{i=1}^{d} \alpha_i \mathrm{E}[x_{t+1-i}] + \mathrm{E}[\varepsilon_{t+1}] + \mu$$

we obtain:

$$m = \frac{\mu}{1 - \sum_{i=1}^{d} \alpha_i}$$

which is well defined due to the stability assumption. There are several well established methods for system identification in linear autoregressive models [31]; in the following, we will use the Yule-Walker method for estimating the parameters of the model. In particular, we define the following correlations:

$$c_k = \mathrm{E}[(x_{i+1} - m)(x_{i+1-k} - m)] = \sum_{j=1}^{d} \alpha_j c_{|j-k|} \qquad \forall k = 1, \ldots, d$$

Introducing the vector $\mathbf{c} = (c_1, c_2, \ldots, c_d)^{\mathrm{T}}$ and the correlation matrix C:

$$C = \begin{pmatrix} c_0 & c_1 & \ldots & c_{d-1} \\ c_1 & c_0 & \ldots & c_{d-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{d-1} & c_{d-2} & \ldots & c_0 \end{pmatrix}$$

we see that $\mathbf{c} = C\boldsymbol{\alpha}$, hence:

$$\boldsymbol{\alpha} = C^{-1}\mathbf{c} \tag{2}$$

provided that C^{-1} exists.

When we observe a time series X comprising n observations, we can estimate $\boldsymbol{\alpha}$ by replacing the elements of c and C in Eq. 2 by the sample correlations which are the unbiased estimators of the true correlations:

$$\hat{c}_k = \frac{1}{n-d}\sum_{i=d}^{n-1}(x_{i+1}-\hat{m})(x_{i+1-k}-\hat{m})$$

where $\hat{m}$ is the mean value of the series. At this point we can pose the problem of estimating the parameters in this way [31]:

$$\hat{\boldsymbol{\alpha}} = \hat{C}^{-1}\hat{\mathbf{c}} \tag{3}$$

Once we have $\hat{\boldsymbol{\alpha}}$, we can estimate the other parameters of the model μ and γ.

$$\hat{\mu} = \hat{m}(1-\sum_{i=1}^{d}\hat{\alpha}_i) \tag{4}$$

Defining the estimated stochastic term as

$$\hat{\varepsilon}_{i+1} = x_{i+1} - \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i - \hat{\mu} \tag{5}$$

we obtain the following expression for the estimated variance of the stochastic term [31]

$$\hat{\gamma}^2 = \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\hat{\varepsilon}_{i+1}\right)^2. \tag{6}$$

B. Information theoretic measure for novelty detection

The main tool that we will consider in this section is the *Kullback-Leibler (KL) divergence* between two distributions $p(x)$ and $q(x)$, also known as *relative entropy*, defined as:

$$\mathrm{KL}\left[p(x)\|q(x)\right] = \int p(x)\log\left[\frac{p(x)}{q(x)}\right]dx. \tag{7}$$

The KL divergence is often characterised as an information theoretic quantity as follows (for a comprehensive overview see, e.g., [29], [32], [33]): consider some unknown distribution $p(x)$, and suppose we wish to use another (simpler) distribution $q(x)$ to build a coding scheme to transmit values of x to a receiver. In the assumption of an efficient coding scheme, the KL divergence measures how many bits per symbol are wasted by using the coding scheme based on $q(x)$ instead of $p(x)$. As perfect compression can be achieved only when $q(x)$ and $p(x)$ coincide almost everywhere, we see that the KL divergence is always positive and provides a measure of the dissimilarity between $p(x)$ and $q(x)$. This can be interpreted as the amount of information about the true distribution which is lost by adopting the approximating distribution $q(x)$. Note, however, that the KL divergence is not a metric, as it is not symmetric and does not satisfy the triangular inequality.

In [4] we proposed a novelty detection method for i.i.d. data based on the idea of measuring the information content of a new data point. To define this information content, we considered the following case: let us assume a parametric generative model for the observations given by a probability distribution $p(x|\theta)$. Let us assume we have a training set X and a point x_* we wish to test for novelty. Let $\hat{\theta}$ be the maximum likelihood estimate of the parameter θ obtained using the training set only, and $\hat{\theta}_*$ be the estimate using the augmented set $\{X, x_*\}$. The information content of the new point x_* was then defined as

$$\mathcal{I}(x_*) = \mathrm{KL}[p(x|\hat{\theta})\|p(x|\hat{\theta}_*)].$$

Remarkably, the distribution of this quantity resulted to be analytically computable in the Gaussian case, yielding a test which is *independent* from the statistics of the generating distribution, and dependent only on the dimensionality and number of available data points. This test turned out to be closely related to the classical F-test, but the new interpretation allowed us to extend this concept to non-trivial cases such as the mixture of Gaussian case.

C. Perturbative measure for novelty detection

In this Section, we introduce the measure for novelty detection we intend to use. We start our discussion from the information theoretic measure for novelty detection based on the KL divergence. Unlike the i.i.d. Gaussian case, the complex dependence of the AR parameters on the data means that the information content cannot be analytically computed. We therefore replace it with a simpler measure which approximates it well in the regions of large deviance, i.e. for points that are likely to be novel.

We consider the effect of the addition of a new point x_* to the training set X on the estimation of the parameters of the AR model, in the null hypothesis that the new point x_* is generated by the same stochastic process as the training set. We denote the updated parameters by $\hat{\boldsymbol{\alpha}}_*$, $\hat{\mu}_*$, and $\hat{\gamma}_*^2$, and the stochastic terms by $\hat{\varepsilon}_{i+1}^*$. Note that to simplify the notation we will use $\varepsilon_{n+1} = \varepsilon_*$ in the sums to denote the stochastic term associated to $x_{n+1} = x_*$.

To estimate the impact of the new point on the identification of the process, we compute the Kullback-Leibler divergence between the distribution of the stochastic term when estimated with and without x_* [29]

$$\mathcal{I}_{\mathrm{KL}}(x_*) = \mathrm{KL}\left[\mathcal{N}(\varepsilon|0,\hat{\gamma}^2)\|\mathcal{N}(\varepsilon|0,\hat{\gamma}_*^2)\right] = \int \mathcal{N}(\varepsilon|0,\hat{\gamma}^2)\log\left[\frac{\mathcal{N}(\varepsilon|0,\hat{\gamma}^2)}{\mathcal{N}(\varepsilon|0,\hat{\gamma}_*^2)}\right] d\varepsilon. \tag{8}$$

Recalling the information-theoretic interpretation of the KL divergence, and remembering that asymptotically ML estimates converge to the true values, we see that one could interpret $\mathcal{I}_{\mathrm{KL}}(x_*)$ as the amount of information about the true process that is gained through the addition of the new data point. Therefore,

we interpret the quantity $\mathcal{I}_{\mathrm{KL}}(x_*)$ as the *information content* of the new point given a training set X, and in the following we will use an approximate expression for it that will be more easy to analyze. Notice that, to alleviate the notation, the dependency on the training set is not indicated explicitly in the following. The rest of the paper is devoted to deriving a tractable approximation of this information content that can be used for statistical testing purposes.

The KL divergence between two Gaussian distribution is readily obtained from its definition and yields:

$$\mathcal{I}_{\mathrm{KL}}(x_*) = \frac{1}{2}\left[\log\left(\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}\right) - 1 + \frac{\hat{\gamma}^2}{\hat{\gamma}_*^2}\right] \tag{9}$$

The information content measured in this way is a function of the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ only. This consideration forms the basis of the proposed perturbative approach. Assuming that the ML estimates for the parameter γ are not dramatically changed by the addition of the new point x_*, we see that the dominant term in equation (9) is a monotonic function of

$$\mathcal{I}(x_*) = \frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}, \tag{10}$$

which is a much easier expression to analyze. It should be noted, however, that we are replacing $\mathcal{I}_{\mathrm{KL}}(x_*)$, a non-monotonic function in the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$, by a monotonic function of it. This is exemplified in Fig. 1, which shows the KL divergence (left panel) and the deviance $\hat{z}^2$ (right) in the i.i.d. case as a function of a new point x_* (the deviance is defined as the squared distance of the new point from the estimated mean $\hat{\mu}$, divided by the estimated variance $\hat{\sigma}^2$ and it is connected to the ratio of the estimated variances $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ [4]). As expected, the KL divergence is minimal when the distance between the new point and the true mean is about one standard deviation. It then rises slightly when the new point is very close to its mean value; this highlights that new points very close to the mean carry as much information as points at about one and a half standard deviation from the mean. We see from Fig. 1, however, that both $\mathcal{I}(x_*)$ and $\mathcal{I}_{\mathrm{KL}}(x_*)$ are monotonic in a large region on the tails of the distribution, corresponding to about 30% of the total area of the Gaussian. Therefore, in realistic novelty detection scenarios, we see that testing the tails of the KL distribution or testing the tails of the distribution of the noise ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ is essentially equivalent.

For this reason, from now on we will focus on $\mathcal{I}(x_*)$ as a measure of information content. Making use of equation (6), we can rewrite this in terms of the estimated stochastic terms as

$$\mathcal{I}(x_*) = \frac{n-d}{n-d+1}\frac{\sum_{i=d}^{n}\left(\hat{\varepsilon}_{i+1}^*\right)^2}{\sum_{i=d}^{n-1}\left(\hat{\varepsilon}_{i+1}\right)^2} \tag{11}$$

When x_* has an additive stochastic term falling in regions of low density of $\mathcal{N}(0,\gamma^2)$, its information content will be unexpectedly high. In general, this method is not able to detect changes in the distribution

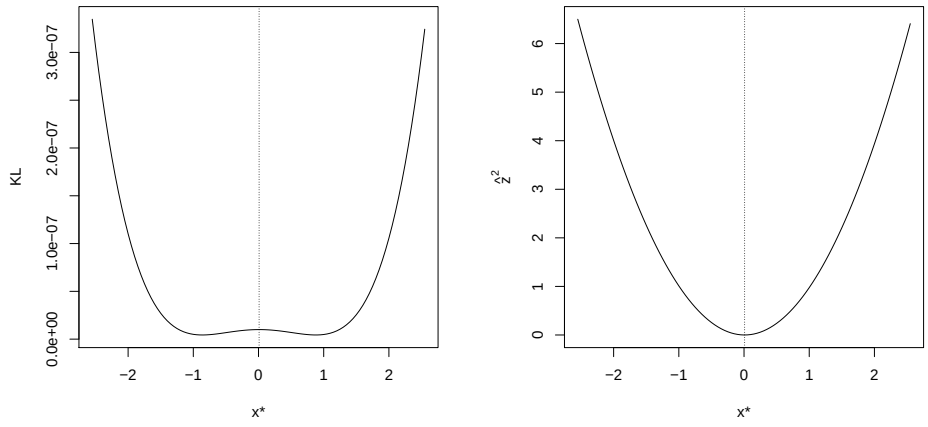

Figure 1. KL divergence (left panel) and deviance $\hat{z}^2 = \frac{(x_* - \hat{\mu}^2)}{\hat{\sigma}^2}$ (right panel) for i.i.d. Gaussian data.

of the process; instead, it is designed to analyze the degree of novelty of the stochastic contribution of x_*. Setting a threshold on the distribution of the information content would allow to flag such situations. The information content measured through the KL divergence will be a distribution with respect to the training set X and the test point x_*. The threshold can be set on the basis of the quantiles of such a distribution when x_* comes from the same model as the data points in X. In this case, in the same spirit of statistical testing, the area of the density function starting from the threshold has to be set on the basis of the percentage of normal data points that we are willing to flag as novel (false positives).

As we will see shortly, the distribution of $\mathcal{I}(x_*)$ with respect to X and x_* is not tractable. We propose an approximation scheme in order to approximate it, leading to a tractable novelty detection method for linear autoregressive time series.

D. Proposed novelty detection method for autoregressive time series

The strategy that we will follow to obtain a tractable test for novelty detection can be summarized in the following steps:

1) expand $\mathcal{I}(x_*)$ in terms of the true values of the parameters;
2) simplify $\mathcal{I}(x_*)$ by using Taylor expansions;
3) neglect random variables with zero expectation and order higher than the second in $1/n$ ($\mathcal{I}(x_*)$ itself is a function in $O(1/n)$);
4) $\mathcal{I}(x_*)$ results in a F-distributed variable with a multiplicative correction term;
5) approximate the multiplicative term by its expectation.

We will present these steps here, and will include few mathematical details in the appendix for clarity.

1) Step 1: We start by writing the estimates of the various parameters as their true values plus a correction term due to the fact that the estimation is based on a finite set of observations; for example, for the linear coefficients, we have

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha} \qquad \hat{\boldsymbol{\alpha}}_* = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_*.$$

Since our test will be essentially based on testing the stochastic term in a new point in the series, we are particularly interested in the estimates of the stochastic terms

$$\hat{\varepsilon}_{i+1} = \varepsilon_{i+1} + \Delta\varepsilon_{i+1} \qquad \hat{\varepsilon}^*_{i+1} = \varepsilon_{i+1} + \Delta\varepsilon^*_{i+1} \tag{12}$$

From the definition of $\hat{\varepsilon}_{i+1}$:

$$\Delta\varepsilon_{i+1} = -\Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \Delta\mu \tag{13}$$

We are interested in making explicit the dependence of the terms in this equation on the training and test data points. By using the model assumptions, we can rewrite the entries $\hat{c}_k$ as

$$\frac{1}{n-d}\sum_{i=d}^{n-1}\left(\sum_{j=1}^{d}\alpha_j x_{i+1-j} + \varepsilon_{i+1} + \mu - \hat{m}\right)(x_{i+1-k} - \hat{m})$$

After some computations, we obtain $\hat{\mathbf{c}} = \hat{C}\boldsymbol{\alpha} + \boldsymbol{\psi}$, where we introduced the vector:

$$\boldsymbol{\psi} = \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \mu\frac{\Delta m}{m}\right)(\mathbf{x}_i - \hat{m}\mathbf{e})$$

where $\mathbf{e} = (1, 1, \ldots, 1)$. If we rewrite the estimate $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}$, we identify $\Delta\boldsymbol{\alpha} = \hat{C}^{-1}\boldsymbol{\psi}$.

Introducing $\delta_{ji} = (\mathbf{x}_j - \hat{m}\mathbf{e})\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e})$, we finally obtain:

$$\Delta\varepsilon_{i+1} = \frac{\mu\Delta m}{m(n-d)}\sum_{j=d}^{n-1}\delta_{ji} - \frac{1}{n-d}\sum_{j=d}^{n-1}\varepsilon_{j+1}\delta_{ji} - \Delta m\frac{\mu}{m} \tag{14}$$

2) Step 2: Substituting the relations (12) into the computations of $\hat{\gamma}^2$ and $\hat{\gamma}^2_*$ allows to show explicitly the dependencies from the stochastic terms ε_{i+1} and their corrections $\Delta\varepsilon_{i+1}$ and $\Delta\varepsilon^*_{i+1}$. Defining:

$$k = \sum_{i=d}^{n-1}(\Delta\varepsilon_{i+1})^2 + 2\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varepsilon_{i+1} \tag{15}$$

$$k^* = \sum_{i=d}^{n}(\Delta\varepsilon^*_{i+1})^2 + 2\sum_{i=d}^{n}\varepsilon_{i+1}\Delta\varepsilon^*_{i+1} \tag{16}$$

we get:

$$\mathcal{I}(x_*) = \frac{n-d}{n-d+1}\frac{\sum_{i=d}^{n}\varepsilon^2_{i+1}}{\sum_{i=d}^{n-1}\varepsilon^2_{i+1}}\left[\frac{1 + \frac{k^*}{\sum_{i=d}^{n}\varepsilon^2_{i+1}}}{1 + \frac{k}{\sum_{i=d}^{n-1}\varepsilon^2_{i+1}}}\right] \tag{17}$$

We compute an approximation of $\mathcal{I}(x_*)$ by using the Taylor expansion of the term in brackets in Eq. 17, and the following Taylor expansion:

$$\frac{1}{\sum_{i=d}^{n} \varepsilon_{i+1}^2} \simeq \frac{1}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \left(1 - \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right) \tag{18}$$

Letting $k^* = k + \Delta k$ and substituting equation (18) into equation (17) we obtain

$$\mathcal{I}(x_*) \simeq \frac{n-d}{n-d+1} \left(1 + \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right) \left[1 + \frac{\Delta k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} - \frac{k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right] \tag{19}$$

In this operation we are neglecting terms having expectation in $O(1/n^3)$; a detailed analysis of the rates of convergence of the various terms in equation (19) is given in the appendix. Also, the same analysis shows that the standard deviation of the various terms decays with the same order in n, hence giving a weak convergence result.

3) Step 3: The expression for the approximate information content given in equation (19) is made up of six terms. The first term is a constant offset term $\frac{n-d}{n-d+1}$. The second term is proportional to the ratio $\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$; since the (true) stochastic terms ε are all independent, the distribution of this term is recognized as an F random variable with 1 and $n-d$ degrees of freedom,

$$\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim \frac{1}{n-d} F_{(1,n-d)}. \tag{20}$$

The expectation $\mathrm{E}\left[\frac{\Delta k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right]$ can be shown to be vanishingly small (see the appendix for details); therefore, we neglect the two terms involving $\frac{\Delta k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$. Finally, the term

$$\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \frac{k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim O\left(\frac{1}{n^3}\right)$$

and is therefore neglected. The remaining term $\frac{k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$ is the main proposed correction and is analyzed in detail in the following step.

4) Step 4: The remaining correction term contains the same F-distributed variable, and is multiplied by a random variable that we will replace by its expectation:

$$\mathcal{I}(x_*) \simeq \frac{n-d}{n-d+1} \left(1 + \tau \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right) \tag{21}$$

with

$$\tau = 1 - \mathrm{E}\left[\frac{k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right] \tag{22}$$

In this step we are neglecting the covariance between the F-distributed variable $\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$ and $\frac{k}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$. In our experimental investigation, this covariance did seem to be indeed negligible, except when the values of the autoregressive parameters $\boldsymbol{\alpha}$ were very close to boundary of the region of stability ($|(1-\boldsymbol{\alpha})| \sim 0$).

Table I

PSEUDO-CODE OF THE PROPOSED NOVELTY DETECTION METHOD FOR AUTOREGRESSIVE TIME SERIES.

1) set the false positive rate ρ;
2) estimate the parameters of an $\mathrm{AR}(d)$ time series on X;
3) compute F_ρ that is the $(1-\rho)$-th quantile of an $F_{(1,n-d)}$;
4) compute the threshold θ_ρ corresponding to the rejection rate ρ:
$$\theta_\rho = \frac{n-d}{n-d+1}\left[1+\frac{1}{n-d}F_\rho\left(1+\frac{d}{n-d}+\frac{1}{n}\right)\right]$$
5) compute the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ given a new observation x_*;
6) **if** $(\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} > \theta_\rho)$ **then** flag $\mathbf{x}_*$ as outlier
7) **else** flag $\mathbf{x}_*$ as normal

5) Step 5: The computation of the expectation τ up to the first order in $1/n$ is straightforward but somewhat intricate, and is given in the appendix. The final result is a simple correction involving solely the size of the training set and the order of the autoregressive model

$$\tau \simeq 1 + \frac{d}{n-d} + \frac{1}{n} \tag{23}$$

After this analysis, we obtain an F-test for the ratio with a correction that depends only on n, and d. Finally, the test we propose is based on:

$$\mathcal{I}(x_*) \simeq \frac{n-d}{n-d+1}\left[1+\frac{1}{n-d}F_{(1,n-d)}\left(1+\frac{d}{n-d}+\frac{1}{n}\right)\right] \tag{24}$$

Fig. 2 illustrates the steps of the approximation of $I(x_*)$ as a function of ε_* for an $\mathrm{AR}(4)$. As we can see, the dark shaded area in the bottom-right plot, that represents the final approximation in Eq. 24, captures quite well the distribution of the true $I(x_*)$ over different resampling of the series (light shaded area). Also, the difference between the approximations in Eq. 24 and Eq. 21 is hardly noticeable, showing that the last step of the approximation is fairly accurate.

Setting a rejection rate, we can easily compute the quantiles of $\mathcal{I}(x_*)$. Such quantiles are a simple combination of the quantiles of an F distribution, n, and d. For large values of n, the correction terms vanish, leaving the F-test only as we would expect, since the estimates of the stochastic terms will converge to the true ones. For small values of n, the correction allows to cope with the fact that the parameter estimation has been performed on a short time series. In Tab. I, we report the steps comprising the novelty detection method for autoregressive time series.

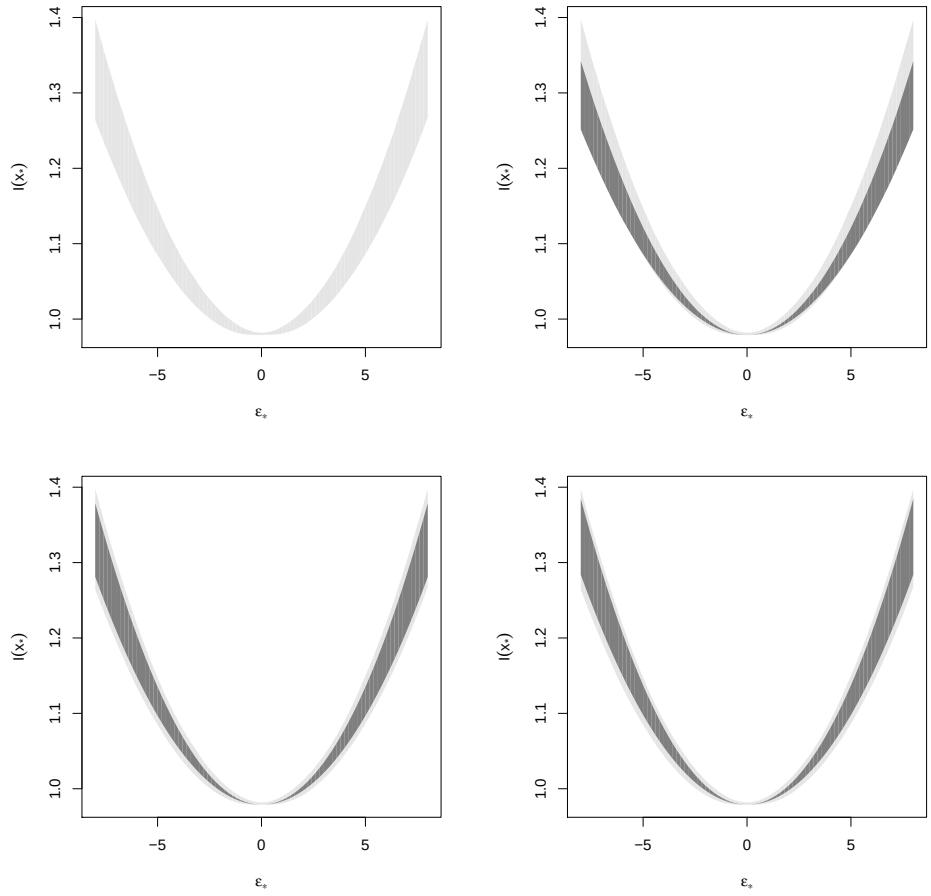

Figure 2. Visualization of the approximation steps of $I(x_*)$ as a function of ε for an $\mathrm{AR}(4)$ with $\alpha = (0.3, -0.2, -0.1, 0.05)$. The shaded areas represent values between the first and third quartile of the distribution over 1000 repetitions where we resample a series of $n = 50$ points. Top-left: the true distribution of $I(x_*)$. Top-right: approximation in Eq. 19 (dark gray) over the true distribution of $I(x_*)$. Bottom-left: approximation in Eq. 21 (dark gray) over the true distribution of $I(x_*)$. Bottom-right: approximation in Eq. 24 (dark gray) over the true distribution of $I(x_*)$.

III. Experimental Results

A. Competing methods

We compare the proposed method, that we will denote as PM (Perturbative Method), against three others that we will call ML (Maximum Likelihood), F-test, and MP (Ma and Perkins [26]).

1) ML: In the ML method, we consider the following residual:

$$\hat{\varepsilon}_* = x_* - \hat{\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{x}_n - \hat{\mu} \tag{25}$$

and treat the estimated parameters as the true ones. In this case, $\hat{\varepsilon}_*$ is compared to the quantiles of $\mathcal{N}(0, \hat{\gamma}^2)$ corresponding to the selected false alarm rate.

2) F-test: The F-test method, instead, is based on the classical statistical F-test, which is the most powerful test for i.i.d. Gaussian data. We will test the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ without the correction term given by the variability in the parameters thus assuming that the estimate of the stochastic terms are distributed as the true ones:

$$\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} \sim \frac{n-d}{n-d+1}\left[1 + \frac{1}{n-d}F_{1,(n-d)}\right] \tag{26}$$

3) MP: The algorithm proposed by Ma and Perkins [26] is based on One-Class SVMs [34] (1-SVM). The methods starts by embedding the time series in a d dimensional space by sliding a window of length d over it. Once the set of $n-d+1$ vectors is obtained, 1-SVM with a Gaussian kernel having variance σ^2 and upper bound on the fraction of outliers ν is applied. This leads to a decision function on whether the vectors fall in a "normal" region or not. When a vector falls in the "outlier" region, all its time points are flagged as outliers. Ma and Perkins suggest to run such procedure for different values of d and flag a time point when it resulted in a novel vector for all the tested dimensions. Given that the time series will have temporal correlations, it has been proposed to project the d-dimensional vectors on the hyperplane having normal vector $\mathbf{e} = (1, 1, \ldots, 1)$. In the experiments, we tested both the unprojected and projected versions; in the experiments, we found that the projected version achieves better performances than the unprojected one in terms of accuracy and false positive rate. Therefore, we will report the results for the projected version only. Unfortunately, [26] does not provide any guidelines on how to set σ and the values of the dimensions. Since the number of possible combinations of such parameters is too large to be explored meaningfully, we decided to set the variance of the kernel automatically; therefore for different dimensions, the kernel will assume different values (in contrast to what shown in [26]). We follow the idea that we want to capture the inner region of the vector representation as the normal region. For this reason, we compute an empirical distribution of the squared pairwise distances among the data vectors, and set σ^2 to correspond to its 95-th percentile. We set the upper bound of outliers in 1-SVM $\nu = \rho$ (ρ is the false positive rate we are willing to tolerate - Tab. I). The values of d for which we test the algorithms are reported for each series throughout the paper.

B. Synthetic data sets

1) Known model order - Test series without novelties: We check the behavior of the proposed method on a set of four synthetically generated linear autoregressive time series with different orders and parameters (see Tab. II).

Table II

PARAMETERS OF THE FOUR SYNTHETIC TIME SERIES. $U_{[-0.1,0.1]}$ STANDS FOR THE UNIFORM DISTRIBUTION IN THE INTERVAL $[-0.1, 0.1]$.

	d	μ	γ	$\boldsymbol{\alpha}$
Synth1	1	2	0.1	(0.3)
Synth2	5	1	0.5	$(0.18, 0.13, 0.12, -0.14, -0.13)$
Synth3	10	-3	0.2	$\alpha_i \sim U_{[-0.1,0.1]}$
Synth4	50	0.5	0.1	$\alpha_i \sim U_{[-0.1,0.1]}$

The goal of this analysis is to see if the proposed method is able to achieve, on average, the expected false alarm rate. We perform this analysis generating a training time series of length n and a test series of length 10^5 drawn from the same model parameters. We use the algorithm in Tab. I for all the test points and we compute the number of points of the test series flagged as novel. Since the test series is generated using the same parameters, the points that are flagged as novel are false positives. In this way, we obtain the false alarm rate for a specific training series of length n. We repeat such procedure for different values of n and we average the false positive rate over 200 repetitions for each value of n; in each repetition we generate a new training series of length n.

We generate the time series using the linear autoregressive model presented in Section II (Eq. 1). In order to sample from a stationary process, we generate d values from a $\mathcal{N}(\mu/(1-\sum_{i=1}^{d}\alpha_i), \gamma^2)$ followed by a burn-in period of 1000 samples. The results on the average false alarm rate are reported in Fig. 3 for the four sets of parameters generating the series for a false alarm rate of 1%. Similar figures are obtained for different rejection rates. In the MP method we selected $d = 2, 3, 4, 5$ for Synth1, $d = 3, 5, 7, 9$ for Synth2, $d = 3, 7, 11, 15$ for Synth3, and $d = 3, 15, 35, 55$ for Synth4.

While all methods except MP reach the desired level of false positive rate for large training sets, PM performs consistently better than all other methods when the size of the training set is small. We notice that the MP method particularly suffers from small sample problems; we hypothesize that this is due to difficulties in selecting the appropriate kernel parameters and the set of dimensions for small training set sizes. Also, notice that MP does not seem to converge to the correct false positive rate for large training sets.

2) Unknown model order - Test series without novelties: In this section we report the result of applying the novelty detection algorithms on the four time series in Tab. II, but we do not assume to know the

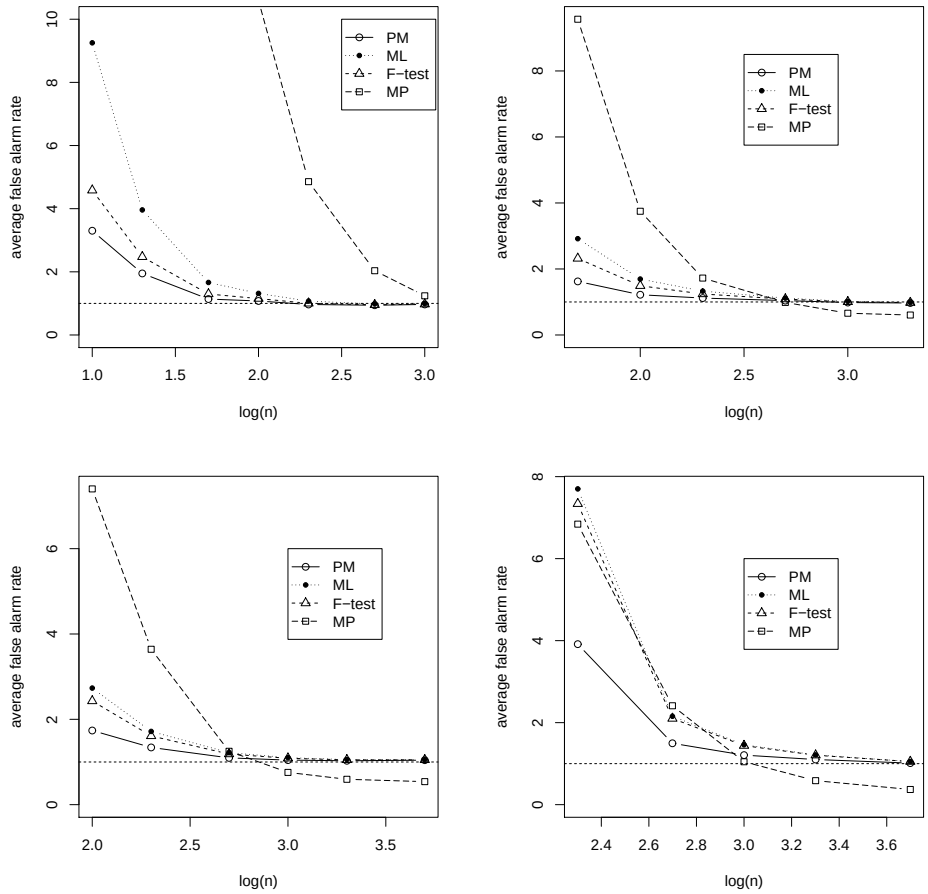

Figure 3. A comparison of the average false alarm rate between the proposed method and the competing ones when the model order is known. The false alarm rate was set to 1%. Top-left: Synth1, top-right: Synth2, bottom-left: Synth3, bottom-right: Synth4.

model order. We repeat the same procedure as in the former case, but every time that we sample a training set of size n, we estimate the model order d. We choose the value of d that maximizes the *Akaike information criterion* (AIC) [35]. The averages of the false alarm rate over 200 repetitions are reported in Fig. 4 for a 1% rejection rate. In the MP method we selected the values of d as in the former section.

3) Known model order - Test series with novelties: In this section we study the behavior of the novelty detection methods on a time series where the test series has been contaminated by some noise. In particular, we generate the test time series by using the same autoregressive model, but letting 5% of the times a stochastic term to be generated from a Gaussian with standard deviation that is 4γ instead

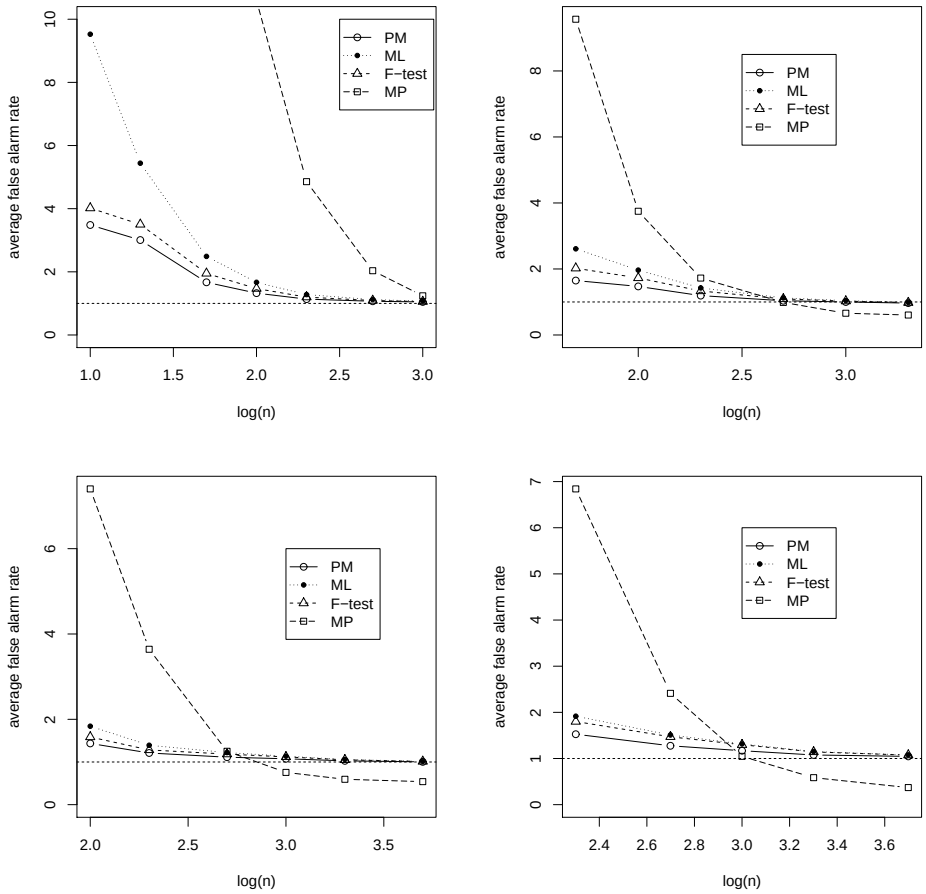

Figure 4. A comparison of the average false alarm rate between the proposed method and the competing ones when the model order is unknown and estimated maximizing the AIC. The false alarm rate was set to 1%. Top-left: Synth1, top-right: Synth2, bottom-left: Synth3, bottom-right: Synth4.

of γ. In Tab. III, we report the true positive rate (TP), false positive rate (FP) and accuracy (ACC) (quartiles over 200 runs) achieved by the tested methods, for different training set sizes and rejection rates. In the MP method we selected the values of d as in the former sections. To assess statistically the difference between the various methods, we performed two-sample t-tests pairwise between the empirical distributions of the FP, TP and ACC results obtained by the various methods on the synthetic data sets (this was not done in the case of Synth1 as the scores were strongly non-Gaussian, making a t-test less meaningful). We see that the proposed method gives significant improvements over all the other methods (at 1% significance) in terms of controlling FP rates (as could be expected, since this was its purpose) but *also in terms of overall accuracy* in all cases. Although we could not perform a statistical test on the

Table III

FALSE POSITIVE RATE (FP), TRUE POSITIVE RATE (TP), AND ACCURACY (ACC) FOR SELECTED VALUES OF n AND REJECTION RATE ρ. IN EACH CELL, WE REPORT THE VALUES OF THE MEDIAN AND IN PARENTHESIS THE FIRST AND THIRD QUARTILE, COMPUTED OVER 200 REPETITIONS. A t-TEST, WITH 1% SIGNIFICANCE, HAS BEEN RUN ON THE DISTRIBUTIONS OF FP, TP, AND ACC WHEN THE CORRESPONDING NORMAL QUANTILE PLOT SHOWED REASONABLE NORMALITY BY VISUAL INSPECTION (ALL EXCEPT THOSE OF SYNTH1); IN THESE CASES THE RESULTS THAT ARE SIGNIFICANTLY BETTER THAN THE OTHERS ARE HIGHLIGHTED IN BOLD.

Synth1	$\rho = 0.01$ $n = 10$		
	FP	TP	ACC
PM	0.010 (0.002, 0.036)	0.451 (0.384, 0.507)	0.962 (0.941, 0.967)
ML	0.054 (0.026, 0.129)	0.584 (0.527, 0.630)	0.928 (0.859, 0.952)
F-test	0.016 (0.005, 0.056)	0.490 (0.425, 0.545)	0.957 (0.924, 0.965)
MP	0.707 (0.529, 0.827)	0.882 (0.788, 0.931)	0.321 (0.211, 0.489)
Synth2	$\rho = 0.05$ $n = 100$		
	FP	TP	ACC
PM	**0.064** (0.049, 0.079)	0.626 (0.608, 0.646)	**0.920** (0.908, 0.935)
ML	0.077 (0.059, 0.092)	**0.642** (0.621, 0.659)	0.910 (0.896, 0.925)
F-test	0.073 (0.056, 0.088)	0.637 (0.617, 0.655)	0.913 (0.899, 0.928)
MP	0.094 (0.075, 0.119)	0.554 (0.514, 0.586)	0.888 (0.866, 0.905)
Synth3	$\rho = 0.01$ $n = 100$		
	FP	TP	ACC
PM	**0.019** (0.012, 0.028)	0.512 (0.488, 0.535)	**0.958** (0.950, 0.963)
ML	0.030 (0.020, 0.041)	**0.542** (0.521, 0.565)	0.949 (0.939, 0.957)
F-test	0.027 (0.018, 0.037)	0.534 (0.513, 0.557)	0.952 (0.942, 0.959)
MP	0.071 (0.054, 0.102)	0.468 (0.416, 0.531)	0.905 (0.880, 0.919)
Synth4	$\rho = 0.05$ $n = 1000$		
	FP	TP	ACC
PM	**0.061** (0.056, 0.066)	0.625 (0.620, 0.630)	**0.924** (0.919, 0.928)
ML	0.068 (0.063, 0.073)	**0.634** (0.629, 0.639)	0.917 (0.912, 0.921)
F-test	0.068 (0.063, 0.073)	**0.634** (0.629, 0.639)	0.917 (0.913, 0.922)
MP	0.100 (0.090, 0.107)	0.501 (0.485, 0.513)	0.881 (0.874, 0.889)

results obtained on Synth1 data set, it is quite clear that the proposed method achieves a better control of the false positive and higher accuracy compared to the other methods.

4) Nonlinear autoregressive time series with non-Gaussian noise: The method that we propose has guarantees to provide reliable results when dealing with linear autoregressive models in a neighborhood

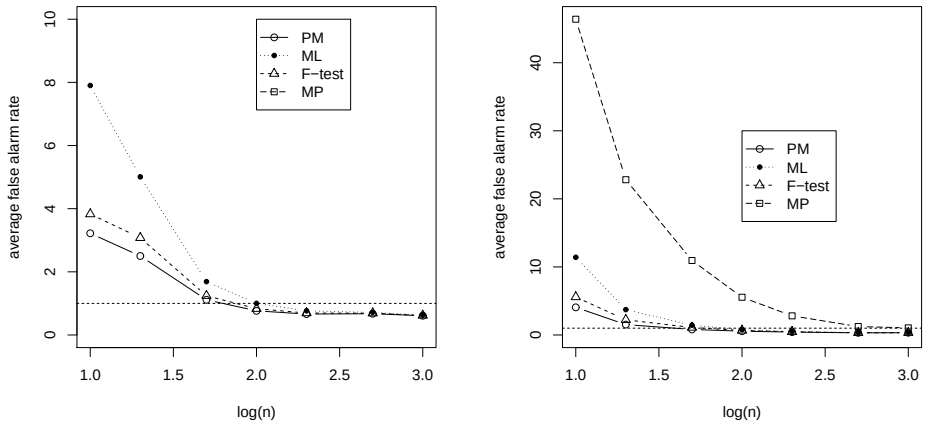

Figure 5. A comparison of the average false alarm rate on two nonlinear non-Gaussian time series (Left: series in Eq. 27 - Right: series in Eq. 28). The false alarm rate was set to 1%. In the left panel, the results from the MP method are not shown since they are very poor.

of the unknown true model order. In this section, we study two scenarios where the models are not linear with non-Gaussian stochastic noise, in order to check how these assumption affect the performances in terms of false positives. The two series we considered are the following:

$$x_{t+1} = -0.1\, x_t + 100\, x_t x_{t-1} + \varepsilon_{t+1} \tag{27}$$

with $\varepsilon_{t+1} \sim 0.4\mathcal{N}(-0.0015, 0.001^2) + 0.6\mathcal{N}(0.001, 0.0005^2)$, and

$$x_{t+1} = \exp(-|x_t + x_t x_{t-1}|) + \varepsilon_{t+1} \tag{28}$$

with $\varepsilon_{t+1} \sim 0.3\mathcal{N}(-0.14, 0.05^2) + 0.7\mathcal{N}(0.06, 0.1^2)$. In both cases it is easy to verify that $E[\varepsilon] = 0$. In Fig. 5 we report the average false positive rate over 200 repetitions for different values of n. Again, we estimate the model order by maximizing the AIC. In the MP method we set $d = 2, 3, 4, 5$ for both the series. It is interesting to notice that for n large, ML, PM, and the F-test converge to a biased solution, where the average false alarm rate is slightly different from the expected one.

C. Real data sets

In this section, we show two applications of the proposed novelty detection method on real time series. The first data set we consider contains the yearly average level (measured in feet) of lake Huron in the northern U.S.A. from 1875 to 1972 (Fig. 6). We train the model on the first 50 years (the vertical solid line in Fig. 6 shows the division between training and the test sets); we estimate the model order by

Figure 6. Novelty detection on the lake Huron data set. The dotted lines are placed on the years identified as novel: 1929, 1931, and 1960.

choosing the model that maximizes the AIC. In this case, we obtained an autoregressive model of order $d = 1$. The proposed method, when we set the rejection rate to 1%, identifies three novelties in the test data (denoted with a black dot), corresponding to the years 1929, 1931, and 1960. The first novelty has been discussed in the geographical literature in Ref. [36]. The other two novelties can be ascribed to unusual precipitation levels in the Michigan area in the previous year[1].

The second data set contains the daily closing value of the FTSE 100 index in the period from January, 1st 2007 to January, 1st 2009. The index is recorded only during working days; therefore, the series contains 253 data points for each year, leading to a series of 506 observations. We trained a linear autoregressive model on the basis of the first 14 months (corresponding to 296 observations); in Fig. 7 the solid line shows the division between training and test sets. We selected a model order $d = 2$ by maximizing the AIC. In the testing phase, we set the rejection rate to 1%; the flagged events are marked by a dot in Fig. 7. We can see that many novel events are detected in the last part of the time series, starting from September 15th 2008, when Lehman Brothers filed for Chapter 11 bankruptcy protection.

IV. Conclusions

In this paper we have introduced a new method for novelty detection in linear autoregressive models with Gaussian noise. The method is based on a perturbative expansion of an information theoretic criterion

[1]Some statistics about the precipitation levels in the area around lake Huron can be found at the following url: http://www.crh.noaa.gov/dtx/cms.php?n=clisum

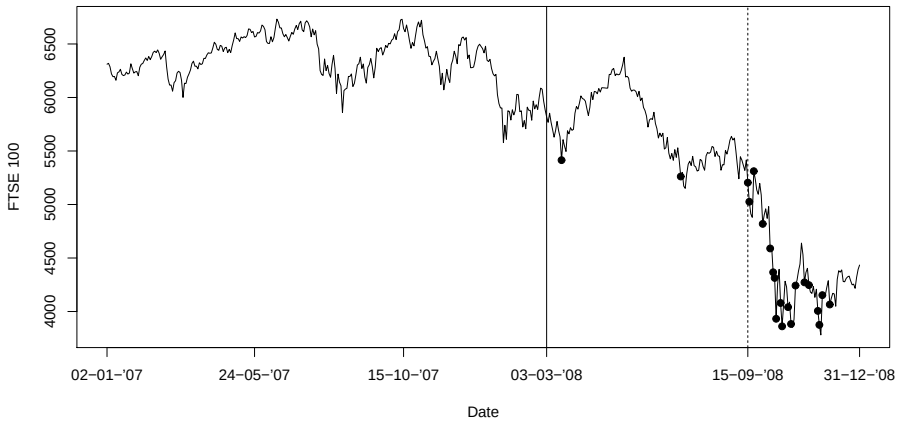

Figure 7. Novelty detection on the FTSE data set. Novelties are marked with a black dot. The dotted line is placed on September 15th 2008 when many observations are identified as novel.

for novelty originally proposed for i.i.d. data in [4]. By further expanding this approximation in powers of $\frac{1}{n}$, where n is the number of samples observed in the training set, we obtain a simple first order correction to the classical F-test, which provides a tight control on the false positives rate for short time series. This correction accounts for the variability in the estimates of the parameters given that they are based on a finite set of observations. Extensive experimentation on synthetic data shows that our approach performs consistently better than competing approaches, with a dramatic difference when the time series is short. At the same time, testing on real data sets shows that the model still capture important novelties which can often be traced back to known events of significance.

Our approach assumes that the modeling and system identification from data is part of a preprocessing separately carried out on the training data. While the approach showed that preprocessing using standard model selection tools is very effective, it would be an interesting area of further research to combine novelty detection with system identification in a single step. Another potential area of interest would be to consider higher order terms, which could lead to significant improvements for short time series.

APPENDIX A
ANALYSIS OF THE APPROXIMATIONS IN THE DERIVATION OF THE PROPOSED NOVELTY DETECTION METHOD

In this section we report the detailed computations involved in the derivation of the correction terms for the proposed novelty detection method. We will start by computing the expectation $\mathrm{E}\left[\frac{k}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right]$

contained in Eq. 22. From the definition of k in Eq. 15, we see that we need to compute the expectations of $\sum_{i=d}^{n-1}(\Delta\varepsilon_{i+1})^2$ and $2\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varepsilon_{i+1}$. Assuming that the estimated stochastic terms and their difference with the true ones are weakly correlated[2], namely $\mathrm{E}[\hat{\varepsilon}_{i+1}\Delta\varepsilon_{i+1}] \simeq 0$, we see that:

$$\mathrm{E}\left[\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq -\mathrm{E}\left[\frac{\sum_{i=d}^{n-1}\Delta\varepsilon_{i+1}^2}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \tag{29}$$

Therefore $\mathrm{E}\left[\frac{k}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq \mathrm{E}\left[\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right]$. This means that in order to estimate $\mathrm{E}\left[\frac{k}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right]$ we need to compute only the expectations of the three terms in $\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}$.

We approximate them as:

$$\mathrm{E}\left[\frac{\mu\Delta m}{m(n-d)}\frac{\sum_{i=d}^{n-1}\sum_{j=d}^{n-1}\varepsilon_{i+1}\delta_{ji}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq 0 \tag{30}$$

$$\mathrm{E}\left[-\frac{1}{n-d}\frac{\sum_{i=d}^{n-1}\sum_{j=d}^{n-1}\varepsilon_{i+1}\varepsilon_{j+1}\delta_{ji}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq -\frac{d}{n-d} \tag{31}$$

$$\mathrm{E}\left[-\frac{\mu}{m}\Delta m\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq -\frac{1}{n} \tag{32}$$

These results are based on few simple considerations and approximations. Terms involving odd powers of ε_{i+1} have expected value zero. We approximate δ_{ij} as the product of two multivariate Gaussian distributed variables in d dimensions with identity covariance matrix; therefore $\mathrm{E}[\delta_{ji}] \simeq 0$ and $\mathrm{E}[\delta_{ji}^2] \simeq d$ for $i \neq j$, and $\mathrm{E}[\delta_{ii}] \simeq d$.

Eq. 31 can be easily obtained by approximating $\sum_{i=d}^{n-1}\sum_{j=d}^{n-1}\varepsilon_{i+1}\varepsilon_{j+1}\delta_{ji}$ inside the expectation by $\sum_{i=d}^{n-1}\varepsilon_{i+1}^2\delta_{ii}$. Since the expectation $\mathrm{E}[\delta_{ii}] \simeq d$ for all i, we can approximate $\sum_{i=d}^{n-1}\varepsilon_{i+1}^2\delta_{ii} \simeq d\sum_{i=d}^{n-1}\varepsilon_{i+1}^2$.

Eq. 32 can be obtained by writing explicitly Δm with respect to the stochastic terms. First, we notice that the expectation contains the multiplication of odd powers in the stochastic terms and the term m contained in $\Delta m = \hat{m} - m$; this expectation will be zero, leaving:

$$\mathrm{E}\left[-\frac{\mu}{m}\Delta m\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq \mathrm{E}\left[-\frac{\mu}{m}\hat{m}\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \tag{33}$$

For the sake of simplicity, we will sketch the procedure to approximate this expectation for the $\mathrm{AR}(1)$ case. We see that the data points can be rewritten as $x_r = \sum_{s=0}^{r-1}\alpha^{r-1-s}\varepsilon_{s+1} + \mathrm{const.}$, simply by using the definition of autoregressive process in Eq. 1. Thus, $\hat{m} = \frac{1}{n}\sum_{r=0}^{n-1}w_{r+1}\varepsilon_{r+1} + \mathrm{const.}$, with $w_{r+1} = \sum_{s=0}^{n-r-1}\alpha^s = \frac{1-\alpha^{n-r}}{1-\alpha}$. The constants will be multiplied by odd powers of the stochastic

[2]This assumption is confirmed by extensive simulations.

terms, and they will have expectation zero. Finally, we approximate $\sum_{i=d}^{n-1}\varepsilon_{i+1}\sum_{r=0}^{n-1}w_{r+1}\varepsilon_{r+1} \simeq \frac{1}{n}\sum_{r=0}^{n-1}w_{r+1}\sum_{i=d}^{n-1}\varepsilon_{i+1}^2$. The mean of the weights results in $\frac{1}{n}\sum_{r=0}^{n-1}w_{r+1} = \frac{1}{1-\alpha}(1-\frac{1}{n}f(\alpha))$, where $f(\alpha)$ is a function of α. Hence:

$$-\frac{\mu}{m}\mathrm{E}\left[\hat{m}\frac{\sum_{i=d}^{n-1}\varepsilon_{i+1}}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq -\frac{\mu}{m}\frac{1}{1-\alpha}\frac{1}{n}\left[1-\frac{1}{n}f(\alpha)\right] \tag{34}$$

Rewriting $\mu/m = 1-\alpha$, and neglecting the order in $O(1/n^2)$ we obtain the expectation in Eq. 32. In a d-order autoregressive model, we can follow the same procedure; neglecting the mixed terms $\alpha_i\alpha_j$, keeping only the coefficients up to the first power, we obtain the general result in Eq. 32.

Finally, it is straightforward to show that:

$$\mathrm{E}\left[\frac{\Delta k}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] = 0 \tag{35}$$

This can be easily obtained by noticing that Δk contains differences of random variables having expectations in $O(1/n)$. In particular:

$$\mathrm{E}\left[\sum_{i=d}^{n}(\Delta\varepsilon_{i+1}^*)^2 - \sum_{i=d}^{n-1}(\Delta\varepsilon_{i+1})^2\right] \simeq (n-d+1)\frac{w_1}{n-d+1} - (n-d)\frac{w_1}{n-d} \simeq 0 \tag{36}$$

$$\mathrm{E}\left[2\sum_{i=d}^{n}\varepsilon_{i+1}\Delta\varepsilon_{i+1}^* - 2\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varepsilon_{i+1}\right] \simeq (n-d+1)\frac{w_2}{n-d+1} - (n-d)\frac{w_2}{n-d} \simeq 0 \tag{37}$$

where w_1 and w_2 are constant values.

ACKNOWLEDGMENT

M. Filippone is currently supported by the Engineering and Physical Sciences Research Council through the grant EP/E052029/1. G. Sanguinetti acknowledges support from the Scottish Government through the Scottish Informatics and Computer Science Alliance (SICSA). This work is independent research commissioned by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Health for Health Research, or the Department of Health.

REFERENCES

[1] C. M. Bishop, "Novelty detection and neural network validation," *IEE Proceedings on Vision, Image and Signal processing*, vol. 141, no. 4, pp. 217–222, 1994.

[2] D. Martinez, "Neural tree density estimation for novelty detection," *IEEE Transactions on Neural Networks*, vol. 9, no. 2, pp. 330–338, Mar 1998.

[3] S. J. Roberts, "Novelty detection using extreme value statistics," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 146, no. 3, pp. 124–129, 1999.

[4] M. Filippone and G. Sanguinetti, "Information theoretic novelty detection," *Pattern Recognition*, vol. 43, no. 3, pp. 805–814, March 2010.

[5] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, August 2004.

[6] B. Schölkopf, R. C. Williamson, A. J. Smola, J. S. Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems 12, NIPS 1999*, S. A. Solla, T. K. Leen, K. R. Müller, S. A. Solla, T. K. Leen, and K. R. Müller, Eds. The MIT Press, 1999, pp. 582–588.

[7] C. Campbell and K. P. Bennett, "A linear programming approach to novelty detection," in *Advances in Neural Information Processing Systems 13, NIPS 2000*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., 2000, pp. 395–401.

[8] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," in *Fourth International Conference on Artificial Neural Networks*, 1995, pp. 442–447.

[9] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," in *Advances in Neural Information Processing Systems 17, NIPS 2004*, December 2004.

[10] V. Barnett and T. Lewis, *Outliers in Statistical Data*, ser. Wiley Series in Probability & Statistics. Wiley, April 1994.

[11] M. Markou and S. Singh, "Novelty detection: a review - part 1: statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[12] J. A. Quinn and C. K. I. Williams, "Known unknowns: Novelty detection in condition monitoring," in *Pattern Recognition and Image Analysis, Third Iberian Conference, IbPRIA 2007*, ser. Lecture Notes in Computer Science, J. Martí, J. M. Benedí, A. M. Mendonça, and J. Serrat, Eds., vol. 4477. Springer, June 2007, pp. 1–6.

[13] C. K. I. Williams, J. A. Quinn, and N. Mcintosh, "Factorial switching kalman filters for condition monitoring in neonatal intensive care," in *Advances in Neural Information Processing Systems 18, NIPS 2005*, December 2005.

[14] C. Archer, T. K. Leen, and A. M. Baptista, "Parameterized novelty detectors for environmental sensor monitoring," in *Advances in Neural Information Processing Systems 16, NIPS 2003*, December 2003.

[15] D. A. Clifton, N. Mcgrogan, L. Tarassenko, D. King, S. King, and P. Anuzis, "Bayesian extreme value statistics for novelty detection in gas-turbine engines," in *Proceedings of IEEE Aerospace*, 2008.

[16] P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," in *Advances in Neural Information Processing Systems 13, NIPS 2000*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2000, pp. 946–952.

[17] Y. Zhan and A. Jardine, "Adaptive autoregressive modeling of non-stationary vibration signals under distinct gear states. part 1: modeling," *Journal of Sound and Vibration*, vol. 286, no. 3, pp. 429–450, September 2005.

[18] K. Choy, "Outlier detection for stationary time series," *Journal of Statistical Planning and Inference*, vol. 99, no. 2, pp. 111–127, 2001.

[19] M. C. Hau and H. Tong, "A practical method for outlier detection in autoregressive time series modelling," *Stochastic Environmental Research and Risk Assessment (SERRA)*, vol. 3, no. 4, pp. 241–260, 1989.

[20] A. Justel, D. Pena, and R. S. Tsay, "Detection of outlier patches in autoregressive time series," *Statistica Sinica*, vol. 11, no. 3, pp. 651–674, 2001.

[21] E. Keogh, S. Lonardi, and Bill, "Finding surprising patterns in a time series database in linear time and space," in *KDD*

'02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2002, pp. 550–556.

[22] H. Louni, "Outlier detection in arma models," *Journal of Time Series Analysis*, vol. 29, no. 6, pp. 1057–1065, November 2008.

[23] A. D. Mcquarrie and C. L. Tsai, "Outlier detections in autoregressive models," *Journal of Computational and Graphical Statistics*, vol. 12, no. 2, pp. 450–471, 2003.

[24] R. P. Adams and D. J. C. Mackay, "Bayesian online changepoint detection," University of Cambridge, Cambridge, UK, Tech. Rep., 2007.

[25] D. Dasgupta and S. Forrest, "Novelty detection in time series data using ideas from immunology," in *In Proceedings of The International Conference on Intelligent Systems*, 1995.

[26] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 613–618.

[27] *Time-series novelty detection using one-class support vector machines*, vol. 3, 2003.

[28] A. L. I. Oliveira, F. B. de Lima Neto, and S. R. de Lemos Meira, "Novelty detection for short time series with neural networks," in *HIS*, ser. Frontiers in Artificial Intelligence and Applications, A. Abraham, M. Köppen, and K. Franke, Eds., vol. 105. IOS Press, 2003, pp. 66–76.

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 99th ed. Wiley-Interscience, August 1991.

[30] S. Eguchi and J. Copas, "Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma," *Journal of Multivariate Analysis*, vol. 97, no. 9, pp. 2034–2040, 2006.

[31] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer, March 2002.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, August 2007.

[33] D. J. C. Mackay, *Information Theory, Inference & Learning Algorithms*, 1st ed. Cambridge University Press, June 2002.

[34] B. Schölkopf, J. C. Platt, J. S. Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[35] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 2003.

[36] M. Jefferson, "Variation in lake huron levels and the chicago drainage canal," *Geographical Review*, vol. 20, no. 1, pp. 133–137, 1930.

Maurizio Filippone Maurizio Filippone received the Master's degree in Physics and the Ph.D. in Computer Science in 2004 and 2008, respectively, both from the University of Genova, Italy. In 2007, he was a Research Scholar with George Mason University, Fairfax, VA. From 2008 to 2009, he was a Research Associate with the University of Sheffield, U.K. He is currently a Research Associate with the inference group, University of Glasgow, U.K. His research interests include computational statistics applied to pattern recognition, bioinformatics, and time-series analysis and forecasting.

Guido Sanguinetti Guido Sanguinetti received his Master's degree in Physics in 1998 from the University of Genova, Italy, and his DPhil in Mathematics in 2002 from the University of Oxford, UK. Between 2004 and 2009, he was a Research Associate and then a Lecturer in Computer Science at the University of Sheffield, UK. He has recently (2010) joined the School of Informatics at the University of Edinburgh, UK, as a SICSA lecturer in Machine Learning. His interests focus on machine learning methodologies with applications in bioinformatics and systems biology, particularly in the field of dynamical systems.