Novelty detection in autoregressive models using information theoretic measures

Maurizio Filippone, Guido Sanguinetti
Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street Sheffield, S1 4DP - United Kingdom.
email - m.filippone@dcs.shef.ac.uk, g.sanguinetti@dcs.shef.ac.uk

July 2009

Technical Report CS-09-06

Abstract

We propose a new method to perform novelty detection in dynamical systems governed by linear autoregressive models. The method extends information theoretic concepts recently introduced for i.i.d. data to the time-series scenario. It is based on a perturbative expansion whose leading term is the classical F-test, and whose $O(\frac{1}{n})$ correction can be computed analytically. We demonstrate on several synthetic examples that the first correction to the F-test can already dramatically improve the control over the false positive rate of the system.

1 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 [4, 19], extreme value statistic [23], information theory [10, 9], support vector methods [24, 5], frequentist [25] and Bayesian [28] non-parametric approaches (for a good review of statistical approaches for novelty detection see *e.g.* [3, 18]).

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 costs of letting some novelties to 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 healthcare [22, 26] to fault detection in engineering [2, 7, 12, 27]. 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 [6, 11, 13, 14, 15, 20]. 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 on-line and based on the model of normality [1, 8, 16, 17, 21]. Another important distinction is 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 recasting the novelty detection problem for temporal data in the framework of information theory, extending our previous work on i.i.d. data [10, 9]. We compute an approximation to the distribution of the information content of a new element of the series by considering the Kullback-Leibler divergence between the estimates of the distributions of the stochastic term. The approximation is carried out by expanding in powers of the inverse of the sample size the estimated parameters with respect to their true values. This procedure yields a modified F-test which is able to accurately control the false positive rate even after a very short training phase.

The paper is organized as follows: in Section 2 we sketch the derivation of the proposed statistical test for novelty detection for linear autoregressive models; in Section 3 we show some experiments on synthetic and real data sets; in Section 4 we draw the conclusions. The full derivation of the method

is reported in the Appendix.

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

Let's 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.

By imposing the 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:

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

In the proposed approach to novelty detection, we use the Yule-Walker method for estimating the parameters of the model. In particular, we define the following expectations:

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

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

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

we see that:

$$\mathbf{c} = C\boldsymbol{\alpha}$$

The inversion of the former equation yields:

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

When we observe a time series X comprising n observations, we can estimate $\boldsymbol{\alpha}$ by using the last equation by substituting c and C with their respective estimates from the series itself:

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

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:

$$\hat{\mathbf{c}} = \hat{C}\hat{\boldsymbol{\alpha}}$$

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

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

$$\hat{\gamma}^2 = \frac{1}{n-d}\sum_{i=d}^{n-1}\left(x_{i+1} - \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i - \hat{\mu}\right)^2$$

Let's now consider the update of the parameters when we add a new data point x_*. We denote such parameters by $\hat{\boldsymbol{\alpha}}_*$, $\hat{\mu}_*$, and $\hat{\gamma}_*^2$. We are interested in evaluating the information content of x_* in the null hypothesis that it has been generated from the same model. In order to do that, we evaluate the Kullback-Leibler divergence between the distribution of the stochastic term when estimated with and without x_*. In particular, we are interested in computing:

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

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

$$\mathrm{KL}\left[\mathcal{N}(\varepsilon|0,\hat{\gamma}^2)\|\mathcal{N}(\varepsilon|0,\hat{\gamma}_*^2)\right] = \frac{1}{2}\left[\log\left(\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}\right) - 1 + \frac{\hat{\gamma}^2}{\hat{\gamma}_*^2}\right] = f\left(\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}\right)$$

For this reason, from now on we will focus on the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ as a measure of information content. If x_* deviates from the normality, its information content will be unexpectedly high. Setting a threshold on the distribution of the information content would allow to flag such situations. The threshold can be set on the basis of the false positive rate that we are willing to tolerate. 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 distribution.

Let's analyze the variance of the stochastic term when we add x_*:

$$\hat{\gamma}_*^2 = \frac{1}{n-d+1}\sum_{i=d}^{n}\left(x_{i+1} - \hat{\boldsymbol{\alpha}}_*^{\mathrm{T}}\mathbf{x}_i - \hat{\mu}_*\right)^2$$

The strategy that we use to obtain the distribution of the ratio is based on the following steps:

1. Write the estimated parameters as their true values plus a term that is given by the fact that the estimation is based on a finite set of observations:

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

$$\hat{\mu} = \mu + \Delta\mu \qquad \hat{\mu}_* = \mu + \Delta\mu_*$$

$$\hat{m} = m + \Delta m \qquad \hat{m}_* = m + \Delta m_*$$

2. Substitute the last relations into the computations of $\hat{\gamma}^2$ and $\hat{\gamma}^2_*$. This allows to show explicitly their dependency from the stochastic terms ε_{i+1} and ε_*.

3. Compute an approximation of the ratio $\frac{\hat{\gamma}^2_*}{\hat{\gamma}^2}$ by using the expansion based on the following leading term:

$$\frac{1}{(n-d)\hat{\gamma}^2} \simeq \frac{1}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$$

   The ratio becomes a function of this form:

$$\frac{\hat{\gamma}^2_*}{\hat{\gamma}^2} \simeq \frac{n-d}{n-d+1}\left[1 + \frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}\right]$$

   where:

$$\Delta = \varepsilon^2_* + \text{correction terms}$$

4. The leading term of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$ is therefore $\frac{\varepsilon^2_*}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$. We know that:

$$\frac{\varepsilon^2_*}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}} \sim \frac{1}{n-d} F_{(1,n-d)}$$

   The leading term of $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$ is $\frac{\varepsilon^2_*}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$; therefore, the leading term of that ratio is distributed as an F. Since Δ contains other correction terms, we decide to approximate the ratio:

$$\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}} \sim \tau \frac{1}{n-d} F_{(1,n-d)}$$

   where τ is a constant that we estimate using the expectation of the correction terms contained in Δ. In practice, we are fitting an $F_{(1,n-d)}$ on the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$. In order to do that, we find the constant τ allowing to match the expected value of the F-distribution with the actual distribution of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon^2_{i+1}}$. We notice that we are keeping the same degrees of freedom of the F-distribution as the leading term.

5. The expectation of the correction terms of Δ, up to the first order in $1/n$ leads to the final result:

$$\tau = 1 + 2\frac{d}{(n-d)} + 2\frac{\mu^2}{m^2}\frac{1}{n} - 2\frac{1}{n}\frac{\mu}{m}$$

Table 1: Pseudo-code of the information theoretic novelty detection method for autoregressive time series.

1. set the false positive rate ρ;
2. estimate the parameters of an 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:

$$\theta_\rho = \frac{n-d}{n-d+1}\left[1+\frac{1}{n-d}F_\rho\left(1+2\frac{d}{(n-d)}+\frac{2}{n}(1-\sum_i\hat{\alpha}_i)^2-\frac{2}{n}(1-\sum_i\hat{\alpha}_i)\right)\right]$$

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

After this analysis, we obtain an F-test for the ratio with a correction that depends on n, d, and $\boldsymbol{\alpha}$ (since $\mu/m = 1-\sum_i \alpha_i$). Since we don't know explicitly $\boldsymbol{\alpha}$, we use its estimate $\hat{\boldsymbol{\alpha}}$. Finally, the test we propose is:

$$\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} = \frac{n-d}{n-d+1}\left[1+\frac{1}{n-d}F_{(1,n-d)}\left(1+2\frac{d}{(n-d)}+\frac{2}{n}(1-\sum_i\hat{\alpha}_i)^2-\frac{2}{n}(1-\sum_i\hat{\alpha}_i)\right)\right]$$

Setting a rejection rate, we can easily compute the quantiles of the ratio in the last equation. Such quantiles are a simple combination of the quantiles of an F distribution and n, d, and the model parameters. For large values of n, the correction terms vanish, leaving the F-test only as we would expect. 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. We report the steps comprising the novelty detection method for autoregressive time series in Tab. 1.

3 Experimental Results

3.1 Synthetic data sets

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. 2).

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 10^6 drawn from the same model parameters. We use the algorithm in

	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	$\alpha_i \sim U_{[-0.1,0.1]}$

Table 2: Parameters of the four synthetic time series. $U_{[-0.1,0.1]}$ stands for the uniform distribution in the interval $[-0.1, 0.1]$.

Tab. 1 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 1000 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 2 (Eq. 1). We first generate d values from a $\mathcal{N}(\mu/(1-\sum_{i=1}^{d}\alpha_i), \gamma^2)$. Then we use the recursive formula 1 starting from these d values. We generate a series of 1000 points longer than n, and we discard the first 1000 values; in this way we expect that the initialization would not affect the generation of the time series.

We compare the proposed method, that we will denote as the KL method (from Kullback-Leibler), against two others that we will call ML (from Maximum Likelihood) and F-test. In the ML method, we consider the following residual:

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

The assumption is that the estimated parameters are 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.

The F-test method, instead, is based on the classical statistical F-test, which is the most powerful test for i.i.d. data. It considers the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$ without the correction term given by the variability in the parameters.

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

Samples of 100 points from the four synthetic time series are shown if Fig. 1. The results on the average false alarm rate are reported in Fig. 2 and 3 for the four sets of parameters generating the series. The two figures correspond to different selection of false positive rates; in Fig. 2 we selected a false alarm rate of 1% whereas in Fig. 3 we set it to 10%.

4 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 an information theoretic criterion for novelty originally proposed for i.i.d. data in [10, 9]. By expanding the KL divergence 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

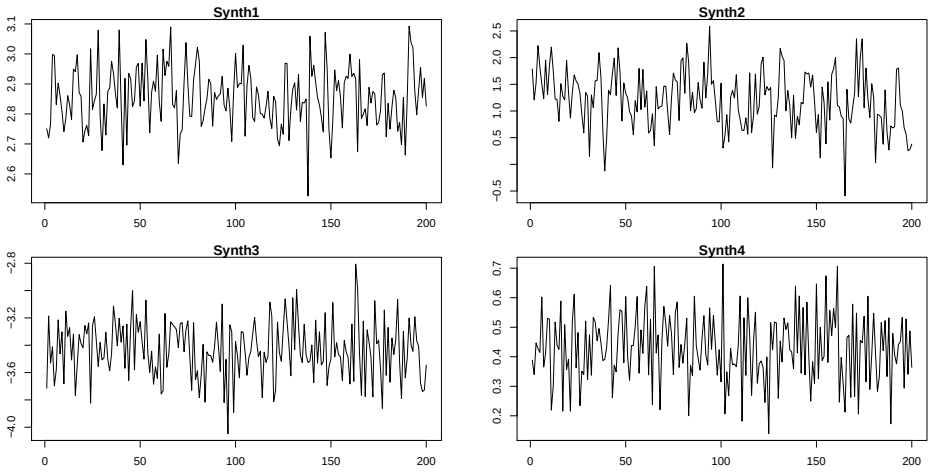

Figure 1: Samples of 100 taken from the four synthetic generated time series.

the classical F-test, which provides a tight control on the false positives rate for short time series. 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.

Our approach assumes that the modeling and system identification from data is part of a pre-processing separately carried out on the training data. In particular, we always tested using the knowledge of the order of the autoregression. While this knowledge was also used for the competing methods, ensuring fairness, 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 corrections in the computation of the KL divergence, which could lead to significant improvements for short time series.

References

[1] R. P. Adams and D. J. C. Mackay. Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK, 2007.

[2] 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.

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

[4] C. M. Bishop. Novelty detection and neural network validation. *IEE Proceedings on Vision, Image and Signal processing*, 141(4):217–222, 1994.

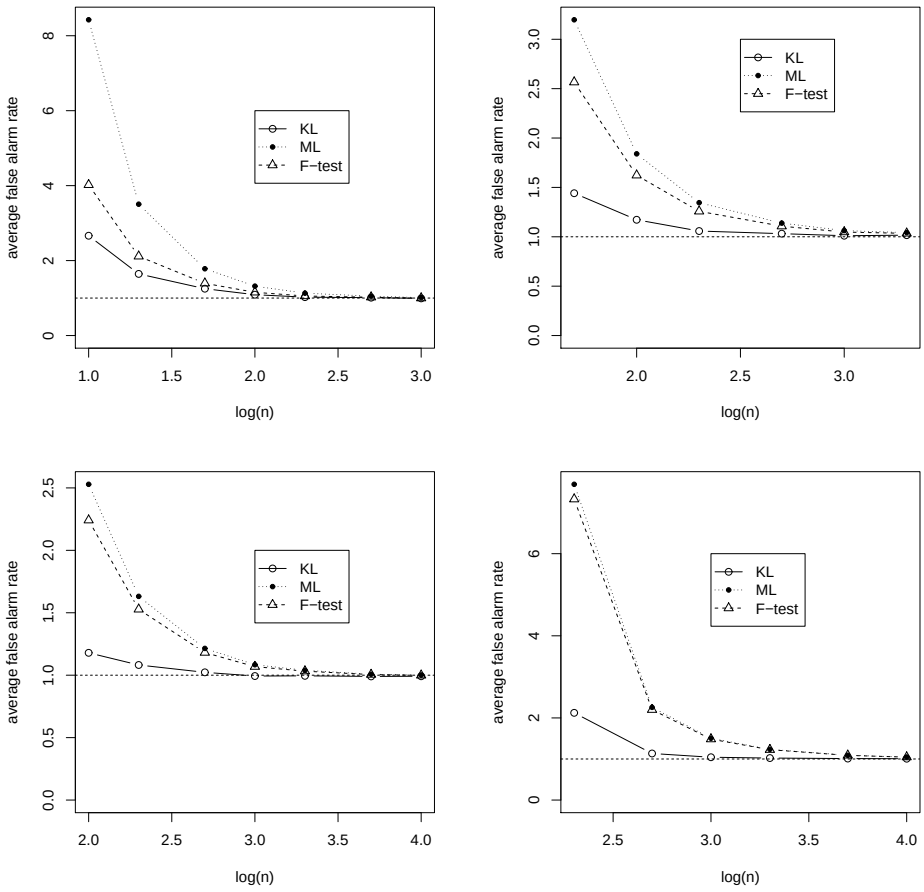

Figure 2: A comparison of the average false alarm rate between the proposed method and the two competing ones. The selected false alarm rate was set to 1%. Top-left: Synth1, top-right: Synth2, bottom-left: Synth3, bottom-right: Synth4.

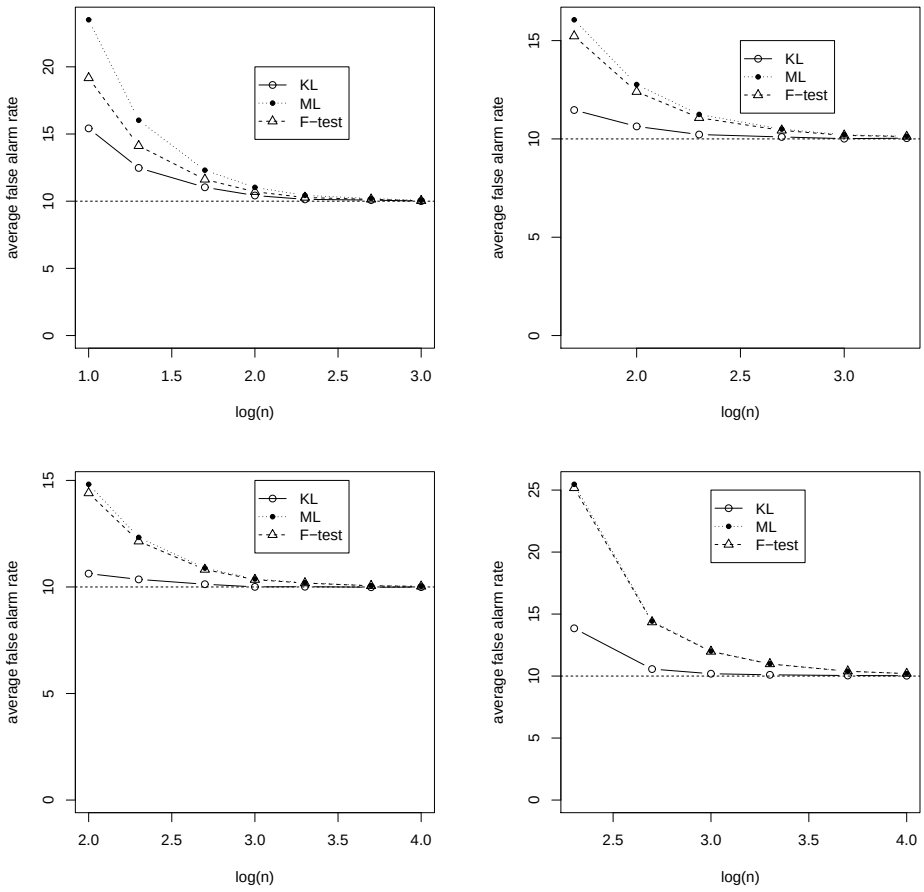

Figure 3: A comparison of the average false alarm rate between the proposed method and the two competing ones. The selected false alarm rate was set to 10%. Top-left: Synth1, top-right: Synth2, bottom-left: Synth3, bottom-right: Synth4.

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

[6] K. Choy. Outlier detection for stationary time series. *Journal of Statistical Planning and Inference*, 99(2):111–127, 2001.

[7] 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.

[8] 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.

[9] M. Filippone and G. Sanguinetti. Information theoretic novelty detection. *Pattern Recognition*. doi:10.1016/j.patcog.2009.07.002

[10] M. Filippone and G. Sanguinetti. Information theoretic novelty detection. Technical Report CS-09-02, Department of Computer Science, University of Sheffield, February 2009.

[11] M. C. Hau and H. Tong. A practical method for outlier detection in autoregressive time series modelling. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 3(4):241–260, 1989.

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

[13] A. Justel, D. Pena, and R. S. Tsay. Detection of outlier patches in autoregressive time series. *Statistica Sinica*, 11(3):651–674, 2001.

[14] 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*, pages 550–556, New York, NY, USA, 2002. ACM.

[15] H. Louni. Outlier detection in arma models. *Journal of Time Series Analysis*, 29(6):1057–1065, November 2008.

[16] 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*, pages 613–618, New York, NY, USA, 2003. ACM.

[17] J. Ma and S. Perkins. Time-series novelty detection using one-class support vector machines. volume 3, pages 1741–1745 vol.3, 2003.

[18] M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[19] D. Martinez. Neural tree density estimation for novelty detection. *IEEE Transactions on Neural Networks*, 9(2):330–338, Mar 1998.

[20] A. D. Mcquarrie and C. L. Tsai. Outlier detections in autoregressive models. *Journal of Computational and Graphical Statistics*, 12(2):450–471, 2003.

[21] A. L. I. Oliveira, F. B. L. Neto, and S. R. L. Meira. Novelty detection for short time series with neural networks. pages 66–75, 2003.

[22] J. A. Quinn and C. K. I. Williams. Known unknowns: Novelty detection in condition monitoring. In J. Martí, J. M. Benedí, A. M. Mendonça, and J. Serrat, editors, *Pattern Recognition and Image Analysis, Third Iberian Conference, IbPRIA 2007*, volume 4477 of *Lecture Notes in Computer Science*, pages 1–6. Springer, June 2007.

[23] S. J. Roberts. Novelty detection using extreme value statistics. *IEE Proceedings on Vision, Image and Signal Processing*, 146(3):124–129, 1999.

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

[25] 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*, pages 442–447, 1995.

[26] 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.

[27] 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*, 286(3):429–450, September 2005.

[28] 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.

A Full derivation of the proposed novelty detection method

In this section we derive explicitly the proposed test for novelty detection. We will start by showing some preliminary material on autoregressive modeling, and then we will present in detail the steps we have been following to obtain the statistical test for linear autoregressive time series.

First of all, let's analyze the following expectations:

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

where we used the facts that $m(\sum_{j=1}^{d} \alpha_j - 1) = -\mu$ and that $\mathrm{E}[\varepsilon_{i+1}(x_{i+1-k} - m)] = 0$. Introducing the vector $\mathbf{c} = (c_1, c_2, \ldots, c_d)^{\mathrm{T}}$ and the matrix C:

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

we can write the following equation relating the expectations and the coefficients of the model:

$$\mathbf{c} = C\boldsymbol{\alpha}$$

Now we present a detailed derivation of the proposed method, following the steps in Section 2.

1. The first step is to write the parameters of the model estimated from data as their true values plus a term that is due to the fact that the estimate is based on a finite set. We will write both the parameters estimated on n and $n+1$ data in such form:

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha} \qquad & \hat{\boldsymbol{\alpha}}_* &= \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_* \\
\hat{\mu} &= \mu + \Delta\mu \qquad & \hat{\mu}_* &= \mu + \Delta\mu_* \\
\hat{m} &= m + \Delta m \qquad & \hat{m}_* &= m + \Delta m_*
\end{aligned}$$

   From data we can estimate the entries of the matrix $\hat{C}$ in the following way:

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

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:

$$\hat{\mathbf{c}} = \hat{C}\hat{\boldsymbol{\alpha}}$$

By using the model assumptions, we can rewrite the entries of $\hat{C}$:

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

After some computations, we obtain:

$$\hat{c}_k = \sum_{j=1}^{d}\alpha_j \hat{c}_{|j-k|}^{(-\min(j,k))} + \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \mu\frac{\Delta m}{m}\right)(x_{i+1-k} - \hat{m})$$

Here we used the notation $\hat{c}_{|j-k|}^{(-\min(j,k))}$ to denote the fact that $\hat{c}_{|j-k|}$ is computed using a batch of point shifted with respect to $\hat{c}_{|j-k|}$.

$$\hat{C} = \begin{pmatrix} c_0^{(-1)} & c_1^{(-1)} & \dots & c_{d-1}^{(-1)} \\ c_1^{(-1)} & c_0^{(-2)} & \dots & c_{d-2}^{(-2)} \\ \vdots & \vdots & \ddots & \vdots \\ c_{d-1}^{(-1)} & c_{d-2}^{(-2)} & \dots & c_0^{(-d)} \end{pmatrix}$$

From the last equation, we see that:

$$\hat{\mathbf{c}} = \hat{C}\boldsymbol{\alpha} + \boldsymbol{\psi}$$

where we introduced the vector:

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

If we rewrite the estimate $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}$, we get:

$$\Delta\boldsymbol{\alpha} = \hat{C}^{-1}\boldsymbol{\psi}$$

We repeat the same procedure when we add a new data point x_*. In this case, we place a star as subscript in the parameter names, and we see that:

$$\hat{\mathbf{c}}_* = \hat{C}_*\hat{\boldsymbol{\alpha}}_* + \boldsymbol{\psi}_*$$

where we introduced the vector:

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

If we rewrite the estimate $\hat{\boldsymbol{\alpha}}_* = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_*$, we get:

$$\Delta\boldsymbol{\alpha}_* = \hat{C}_*^{-1}\boldsymbol{\psi}_*$$

It is convenient now to write $\boldsymbol{\psi}_*$ and $\hat{C}_*$ in an incremental fashion.
Before doing that let's write the incremental forms for the parameters μ and m.

$$\hat{\mu} = \hat{m}(1 - \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{e}) = (m + \Delta m)(1 - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e} - \Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e})$$

$$\hat{\mu} = \mu - m\Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e} + \Delta m\left[1 - (\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{e}\right] = \mu + \Delta\mu$$

We define:

$$\hat{m}_* = \hat{m} + \delta m$$

We notice that:

$$\delta m = \frac{x_* - \hat{m}}{n+1} = \frac{1}{n+1}(\varepsilon_* + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n + \mu - \hat{m})$$

and

$$\Delta m - \Delta m_* = \hat{m} - m - \hat{m}_* + m = -\delta m$$

Going back to the analysis of $\boldsymbol{\alpha}$, we consider:

$$\boldsymbol{\psi}_* = \boldsymbol{\psi} + \Delta\boldsymbol{\psi}$$

Rewriting $\boldsymbol{\psi}_*$, we obtain:

$$\begin{aligned}
\boldsymbol{\psi}_* &= \frac{1}{n-d+1}\left[\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \mu\frac{\Delta m}{m} - \mu\frac{\delta m}{m}\right)(\mathbf{x}_i - \hat{m}\mathbf{e} - \delta m\mathbf{e}) + \right. \\
&\left.\left(\varepsilon_* - \mu\frac{\Delta m}{m} - \mu\frac{\delta m}{m}\right)(\mathbf{x}_n - \hat{m}\mathbf{e} - \delta m\mathbf{e})\right] \\
&= \frac{1}{n-d+1}\left[(n-d)\boldsymbol{\psi} - \delta m\mathbf{e}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \mu\frac{\Delta m}{m}\right) - \mu\frac{\delta m}{m}\sum_{i=d}^{n-1}(\mathbf{x}_i - \hat{m}_*\mathbf{e}) + \right. \\
&\left.\left(\varepsilon_* - \mu\frac{\Delta m_*}{m}\right)(\mathbf{x}_n - \hat{m}_*\mathbf{e})\right]
\end{aligned}$$

Therefore:

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

Concerning $\hat{C}_*$, we analyze its entries:

$$(\hat{c}^*_{|j-k|})^{(-r)} = \frac{1}{n-d+1}\sum_{i=d}^{n}(x_{i+1-j+1-r} - \hat{m}_*)(x_{i+1-k+1-r} - \hat{m}_*)$$

where $r = \min(j,k)$ and $j,k = 1,\dots,d$.

$$
\begin{aligned}
(\hat{c}^*_{|j-k|})^{(-r)} =& \frac{1}{n-d+1}\sum_{i=d}^{n}(x_{i+1-j+1-r} - \hat{m} - \delta m)(x_{i+1-k+1-r} - \hat{m} - \delta m) \\
&= \frac{1}{n-d+1}\left(\sum_{i=d}^{n-1}(x_{i+1-j+1-r} - \hat{m})(x_{i+1-k+1-r} - \hat{m}) - \delta m \sum_{i=d}^{n}(x_{i+1-j+1-r} - \hat{m})\right. \\
&\left. -\delta m \sum_{i=d}^{n}(x_{i+1-k+1-r} - \hat{m}) + \sum_{i=d}^{n}\delta m^2 + (x_{n+1-j+1-r} - \hat{m}_*)(x_{n+1-k+1-r} - \hat{m}_*)\right) \\
&= \frac{n-d}{n-d+1}\hat{c}_k + \frac{n-d}{n-d+1}\delta m^2 - \frac{\delta m}{n-d+1}\left(\sum_{i=d}^{n-1}(x_{i+1-j+1-r} + x_{i+1-k+1-r} - 2\hat{m})\right) \\
&+ \frac{1}{n-d+1}(x_{n+1-j+1-r} - \hat{m}_*)(x_{n+1-k+1-r} - \hat{m}_*)
\end{aligned}
$$

Finally, we get the following expression:

$$\hat{C}_* = \frac{n-d}{n-d+1}\hat{C} + \Delta\hat{C}$$

where:

$$
\begin{aligned}
\Delta\hat{C} &= \frac{n-d}{n-d+1}\delta m^2 \mathbf{e}\mathbf{e}^{\mathrm{T}} - \frac{\delta m}{n-d+1}\left(\sum_{i=d}^{n-1}(\mathbf{e}\mathbf{x}_i^{\mathrm{T}} + \mathbf{x}_i\mathbf{e}^{\mathrm{T}} - 2\hat{m}\mathbf{e}\mathbf{e}^{\mathrm{T}})\right) + \frac{1}{n-d+1}(\mathbf{x}_n - \hat{m}_*\mathbf{e})(\mathbf{x}_n - \hat{m}_*\mathbf{e})^{\mathrm{T}} \\
&= \frac{n-d}{n-d+1}\delta m^2 \mathbf{e}\mathbf{e}^{\mathrm{T}} - \frac{\delta m}{n-d+1}Z + \frac{1}{n-d+1}(\mathbf{x}_n - \hat{m}_*\mathbf{e})(\mathbf{x}_n - \hat{m}_*\mathbf{e})^{\mathrm{T}} \\
&= \frac{1}{n-d+1}\left[(n-d)\delta m^2 \mathbf{e}\mathbf{e}^{\mathrm{T}} - \delta m Z + (\mathbf{x}_n - \hat{m}_*\mathbf{e})(\mathbf{x}_n - \hat{m}_*\mathbf{e})^{\mathrm{T}}\right] \\
&= \frac{1}{n-d+1}A
\end{aligned}
$$

where the matrices A and Z have been introduced for the sake of convenience and their definition is clear from the derivation. Using a Woodbury matrix identity, we see that $\hat{C}_*$:

$$\hat{C}_*^{-1} = \frac{n-d+1}{n-d}\hat{C}^{-1} - \frac{(n-d+1)}{(n-d)^2}\hat{C}^{-1}\left(A^{-1} + \frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1}$$

With abuse of notation, we write:

$$\hat{C}_*^{-1} = \hat{C}^{-1} + \Delta\hat{C}^{-1}$$

where we denoted with $\Delta\hat{C}^{-1}$ the correction of the inverse of $\hat{C}$ to obtain the inverse of $\hat{C}_*$:

$$\Delta\hat{C}^{-1} = -\frac{(n-d+1)}{(n-d)^2}\hat{C}^{-1}\left(A^{-1} + \frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1} + \frac{1}{n-d}\hat{C}^{-1}$$

2. Now we can go back to the analysis of the variance of the stochastic term. We start from its definition and we substitute the estimates of the parameters by the true values plus the correction terms:

$$\begin{aligned}
\hat{\gamma}^2 &= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(x_{i+1} - \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i - \hat{\mu}\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(x_{i+1} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \mu - \Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \Delta\mu\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i - \Delta\mu\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i + m\Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e} - \Delta m\left[1 - (\boldsymbol{\alpha}+\Delta\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{e}\right]\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i + m\Delta\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e} - \Delta m + \Delta m(\boldsymbol{\alpha}+\Delta\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{e}\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \Delta\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{x}_i - \hat{m}) - \Delta m(1 - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{e})\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}) - \Delta m\frac{\mu}{m}\right)^2 \\
&= \frac{1}{n-d}\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \varphi_i\right)^2
\end{aligned}$$

For convenience we introduced:

$$\varphi_i = \boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}) + \Delta m\frac{\mu}{m}$$

Following the same derivation, we rewrite the variance when we consider $n+1$ points:

$$\begin{aligned}
\hat{\gamma}_*^2 &= \frac{1}{n-d+1}\sum_{i=d}^{n}\left(x_{i+1} - \hat{\boldsymbol{\alpha}}_*^{\mathrm{T}}\mathbf{x}_i - \hat{\mu}_*\right)^2 \\
&= \frac{1}{n-d+1}\left[\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \boldsymbol{\psi}_*^{\mathrm{T}}\hat{C}_*^{-1}(\mathbf{x}_i - \hat{m}_*) - \Delta m_*\frac{\mu}{m}\right)^2 + \left(\varepsilon_* - \boldsymbol{\psi}_*^{\mathrm{T}}\hat{C}_*^{-1}(\mathbf{x}_n - \hat{m}_*) - \Delta m_*\frac{\mu}{m}\right)^2\right] \\
&= \frac{1}{n-d+1}\left[\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \varphi_i^*\right)^2 + \left(\varepsilon_* - \varphi_n^*\right)^2\right]
\end{aligned}$$

3. As we saw in Section 2 the information content of a new data point is a function of the ratio:

$$\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} = \frac{n-d}{n-d+1}\frac{\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \varphi_i^*\right)^2 + \left(\varepsilon_* - \varphi_n^*\right)^2}{\sum_{i=d}^{n-1}\left(\varepsilon_{i+1} - \varphi_i\right)^2}$$

Now we write φ_i^* in an incremental fashion $\varphi_i^* = \varphi_i + \Delta\varphi_i$, obtaining:

$$\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} = \frac{n-d}{n-d+1}\left[1 + \frac{\Delta}{(n-d)\hat{\gamma}^2}\right]$$

where we defined Δ as:

$$\Delta = \varepsilon_*^2 + \varphi_n^2 + \sum_{i=d}^{n} \Delta\varphi_i^2 + 2\sum_{i=d}^{n} \varphi_i\,\Delta\varphi_i - 2\sum_{i=d}^{n-1} \varepsilon_{i+1}\Delta\varphi_i - 2\varepsilon_*\varphi_n - 2\varepsilon_*\Delta\varphi_n$$

In order to obtain a tractable test for novelty detection, we decide to approximate the ratio $\frac{\Delta}{(n-d)\hat{\gamma}^2}$. Analyzing the denominator $(n-d)\hat{\gamma}^2$, we see that:

$$(n-d)\hat{\gamma}^2 = \sum_{i=d}^{n-1} (\varepsilon_{i+1} - \varphi_i)^2$$

$$(n-d)\hat{\gamma}^2 = \sum_{i=d}^{n-1} \varepsilon_{i+1}^2 + \sum_{i=d}^{n-1} \varphi_i^2 - 2\sum_{i=d}^{n-1} \varepsilon_{i+1}\varphi_i$$

$$a = \sum_{i=d}^{n-1} \varepsilon_{i+1}^2 \qquad b = \sum_{i=d}^{n-1} \varphi_i^2 - 2\sum_{i=d}^{n-1} \varepsilon_{i+1}\varphi_i$$

The quantity a represents the main contribution to $(n-d)\hat{\gamma}^2$, while b is small when n is large. Therefore, we can use a Taylor expansion of the following ratio (for b small):

$$\frac{\Delta}{a+b} \simeq \frac{1}{a}\Delta - \frac{b}{a^2}\Delta + \dots$$

For the moment we will use only the leading term Δ/a, thus obtaining:

$$\frac{\Delta}{(n-d)\hat{\gamma}^2} \simeq \frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} + \frac{\varphi_n^2 + \sum_{i=d}^{n} \Delta\varphi_i^2 + 2\sum_{i=d}^{n} \varphi_i\,\Delta\varphi_i - 2\sum_{i=d}^{n-1} \varepsilon_{i+1}\Delta\varphi_i - 2\varepsilon_*\varphi_n - 2\varepsilon_*\Delta\varphi_n}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$$

4. The leading term of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$ is therefore $\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$. We know that:

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

We approximate the ratio:

$$\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim \tau \frac{1}{n-d} F_{(1,n-d)}$$

where τ is a correction term that we estimate using the expectation of Δ. In practice, we want to find the constant τ allowing to match the expected value of the F-distribution with the actual

distribution of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}$. We notice that we keep the same degrees of freedom of the F-distribution as the leading term. Computing the expectation of both the sides, we obtain:

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

given that the expected value of an F distributed variable with degrees of freedom 1 and $n-d$ is $(n-d)/(n-d-2)$. The constant τ can be computed as:

$$\tau = (n-d-2)\mathrm{E}\left[\frac{\Delta}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right]$$

Given the form of Δ, we see that:

$$\tau = 1 + z$$

where:

$$z = (n-d-2)\mathrm{E}\left[\frac{\varphi_n^2 + \sum_{i=d}^{n}\Delta\varphi_i^2 + 2\sum_{i=d}^{n}\varphi_i\,\Delta\varphi_i - 2\sum_{i=d}^{n-1}\varepsilon_{i+1}\Delta\varphi_i - 2\varepsilon_*\varphi_n - 2\varepsilon_*\Delta\varphi_n}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right]$$

5. Now we compute an approximation of the expectations in z retaining only the terms up to the first order in $1/n$. Before doing that, we introduce a notation that will help us to keep the notation uncluttered:

$$\delta = \mathbf{e}^{\mathrm{T}}\hat{C}^{-1}\mathbf{e}$$
$$\delta_i = \mathbf{e}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e})$$
$$\delta_{ji} = (\mathbf{x}_j - \hat{m}\mathbf{e})\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e})$$
$$\xi = \mathbf{e}^{\mathrm{T}}\hat{C}^{-1}\left(A^{-1} + \frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1}\mathbf{e}$$
$$\xi_i = \mathbf{e}^{\mathrm{T}}\hat{C}^{-1}\left(A^{-1} + \frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e})$$
$$\xi_{ji} = (\mathbf{x}_j - \hat{m}\mathbf{e})^{\mathrm{T}}\hat{C}^{-1}\left(A^{-1} + \frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e})$$

The variables φ_i can be written as:

$$\begin{aligned}\varphi_i &= \boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e}) + \Delta m\frac{\mu}{m}\\ &= \frac{1}{n-d}\sum_{j=d}^{n-1}\left(\varepsilon_{j+1} - \mu\frac{\Delta m}{m}\right)(\mathbf{x}_j - \hat{m}\mathbf{e})^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i - \hat{m}\mathbf{e}) + \Delta m\frac{\mu}{m}\\ &= \frac{1}{n-d}\sum_{j=d}^{n-1}\left(\varepsilon_{j+1} - \mu\frac{\Delta m}{m}\right)\delta_{ji} + \Delta m\frac{\mu}{m}\end{aligned}$$

The variables φ_i^* can be written as:

$$\varphi_i^* = \boldsymbol{\psi}_*^{\mathrm{T}} \hat{C}_*^{-1}(\mathbf{x}_i - \hat{m}_*) + \Delta m_* \frac{\mu}{m}$$

We are interested in an incremental form for φ_i^*:

$$\varphi_i^* = \varphi_i + \Delta\varphi_i$$

$$\varphi_i^* = (\boldsymbol{\psi} + \Delta\boldsymbol{\psi})^{\mathrm{T}} \left(\hat{C}^{-1} + \Delta\hat{C}^{-1}\right)(\mathbf{x}_i - \hat{m}\mathbf{e} - \delta m\mathbf{e}) + \Delta m_* \frac{\mu}{m}$$

The increments $\Delta\varphi_i$ read:

$$\begin{aligned}\Delta\varphi_i = &- \delta m \boldsymbol{\psi}^{\mathrm{T}} \hat{C}^{-1}\mathbf{e} + \boldsymbol{\psi}^{\mathrm{T}} \Delta\hat{C}^{-1}(\mathbf{x}_i - \hat{m}_*\mathbf{e}) + \Delta\boldsymbol{\psi}^{\mathrm{T}} \hat{C}^{-1}(\mathbf{x}_i - \hat{m}_*\mathbf{e}) \\ &+ \Delta\boldsymbol{\psi}^{\mathrm{T}} \Delta\hat{C}^{-1}(\mathbf{x}_i - \hat{m}_*\mathbf{e}) + \delta m \frac{\mu}{m}\end{aligned}$$

Let's compute each term of $\Delta\varphi_i$.

$$\begin{aligned}-\delta m \boldsymbol{\psi}^{\mathrm{T}} \hat{C}^{-1}\mathbf{e} &= -\delta m \frac{1}{n-d} \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\delta_r \\ &= -\frac{\varepsilon_*}{(n-d)(n+1)} \left[\sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\delta_r\right] - \frac{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n + \mu - \hat{m}}{(n-d)(n+1)} \left[\sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\delta_r\right]\end{aligned}$$

$$\begin{aligned}\boldsymbol{\psi}^{\mathrm{T}} \Delta\hat{C}^{-1}(\mathbf{x}_i - \hat{m}_*\mathbf{e}) = &- \frac{n-d+1}{(n-d)^3} \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)(\xi_{ri} - \delta m \xi_r) \\ &+ \frac{1}{(n-d)^2} \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)(\delta_{ri} - \delta m \delta_r) \\ &= -\frac{n-d+1}{(n-d)^3} \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\xi_{ri} \\ &+ \frac{n-d+1}{(n-d)^3(n+1)} \left[\varepsilon_* \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\xi_r \right. \\ &\left. + (\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n + \mu - \hat{m}) \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\xi_r\right] \\ &+ \frac{1}{(n-d)^2} \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\delta_{ri} \\ &- \frac{1}{(n-d)^2(n+1)} \left[\varepsilon_* \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\delta_r \right. \\ &\left. + (\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n + \mu - \hat{m}) \sum_{r=d}^{n-1} \left(\varepsilon_{r+1} - \mu\frac{\Delta m}{m}\right)\delta_r\right]\end{aligned}$$

$$
\begin{aligned}
\Delta\boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i-\hat{m}_*\mathbf{e}) =& \frac{1}{n-d+1}\left[-\delta m\sum_{r=d}^{n-1}\left(\varepsilon_{r+1}-\mu\frac{\Delta m}{m}\right)(\delta_i-\delta m\delta)\right.\\
&-\mu\frac{\delta m}{m}\sum_{r=d}^{n-1}(\delta_{ri}-\delta m(\delta_i+\delta_r)+\delta m^2\delta)\\
&+\left(\varepsilon_*-\mu\frac{\Delta m_*}{m}\right)(\delta_{ni}-\delta m(\delta_n+\delta_i)+\delta m^2\delta)\\
&\left.-\boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i-\hat{m}_*\mathbf{e})\right]\\
&=-\frac{\varepsilon_*+\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n+\mu-\hat{m}}{(n-d+1)(n+1)}\left[\sum_{r=d}^{n-1}\left(\varepsilon_{r+1}-\mu\frac{\Delta m}{m}\right)(\delta_i-\delta m\delta)\right]\\
&-\frac{(\varepsilon_*+\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_n+\mu-\hat{m})}{(n-d+1)(n+1)}\frac{\mu}{m}\left[\sum_{r=d}^{n-1}(\delta_{ri}-\delta m(\delta_i+\delta_r)+\delta m^2\delta)\right]\\
&+\frac{1}{n-d+1}\left[\left(\varepsilon_*-\mu\frac{\Delta m_*}{m}\right)(\delta_{ni}-\delta m(\delta_n+\delta_i)+\delta m^2\delta)\right]\\
&-\frac{1}{n-d+1}\left[\boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}(\mathbf{x}_i-\hat{m}_*\mathbf{e})\right]
\end{aligned}
$$

$$
\begin{aligned}
\Delta\boldsymbol{\psi}^{\mathrm{T}}\Delta\hat{C}^{-1}(\mathbf{x}_i-\hat{m}_*\mathbf{e}) =& \frac{1}{(n-d)^2}\delta m\sum_{r=d}^{n-1}\left(\varepsilon_{r+1}-\mu\frac{\Delta m}{m}\right)(\xi_i-\delta m\xi)\\
&+\frac{1}{(n-d)^2}\mu\frac{\delta m}{m}\sum_{r=d}^{n-1}(\xi_{ri}-\delta m(\xi_r+\xi_i)+\delta m^2\xi)\\
&-\frac{1}{(n-d)^2}\left(\varepsilon_*-\mu\frac{\Delta m_*}{m}\right)(\xi_{ni}-\delta m(\xi_i+\xi_n)+\delta m^2\xi)\\
&+\frac{1}{(n-d)^2}\boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}\left(A^{-1}+\frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1}(\mathbf{x}_i-\hat{m}_*\mathbf{e})\\
&-\frac{1}{(n-d)(n-d+1)}\delta m\sum_{r=d}^{n-1}\left(\varepsilon_{r+1}-\mu\frac{\Delta m}{m}\right)(\delta_i-\delta m\delta)\\
&-\frac{1}{(n-d)(n-d+1)}\mu\frac{\delta m}{m}\sum_{r=d}^{n-1}(\delta_{ri}-\delta m(\delta_r+\delta_i)+\delta m^2\delta)\\
&+\frac{1}{(n-d)(n-d+1)}\left(\varepsilon_*-\mu\frac{\Delta m_*}{m}\right)(\delta_{ni}-\delta m(\delta_i+\delta_n)+\delta m^2\delta)\\
&-\frac{1}{(n-d)(n-d+1)}\boldsymbol{\psi}^{\mathrm{T}}\hat{C}^{-1}\left(A^{-1}+\frac{\hat{C}^{-1}}{n-d}\right)^{-1}\hat{C}^{-1}(\mathbf{x}_i-\hat{m}_*\mathbf{e})
\end{aligned}
$$

The last term is:

$$\delta m \frac{\mu}{m} = \frac{x_* - \hat{m}}{n+1} \frac{\mu}{m}$$

We see that:

$$\delta m \frac{\mu}{m} = \frac{\varepsilon_*}{n+1} \frac{\mu}{m} + \frac{1}{n+1} \frac{\mu}{m} (\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{x}_n + \mu - \hat{m})$$

Now we have everything we need to compute an approximation of z by substituting all the variables with their expectations. Since many terms involve odd powers of ε_* or ε_{i+1} that have expected value zero, we simply ignore them. We approximate δ_{ij} as the product of two multivariate Gaussian distributed variables in d dimensions with identity covariance matrix:

$$\mathrm{E}[\delta_{ji}] \simeq 0 \qquad \mathrm{E}[\delta_{ji}^2] \simeq d \quad i \neq j$$

$$\mathrm{E}[\delta_{ii}] \simeq d \qquad \mathrm{E}[\delta_{ii}^2] \simeq d(d+2)$$

$$\mathrm{E}[\sum_r \delta_r] \simeq 0 \qquad \mathrm{E}[\sum_r \xi_r] \simeq 0$$

Also, we make the following approximations:

$$\mathrm{E}[\delta m^2] \simeq \frac{\mathrm{var}(x)}{(n+1)^2} \simeq \frac{\gamma^2}{(n+1)^2}$$

$$\mathrm{E}[\Delta m^2] \simeq \frac{\mathrm{var}(x)}{n} \simeq \frac{\gamma^2}{n}$$

$$\mathrm{E}\left[\frac{1}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right] \simeq \frac{1}{(n-d-2)\gamma^2}$$

since the inverse of the sum of $n-d$ squares of Gaussian variables is distributed as an inverse chi square with expectation $1/(n-d-2)$.

We recall that:

$$z = (n-d-2)\mathrm{E}\left[\frac{\varphi_n^2 + \sum_{i=d}^n \Delta\varphi_i^2 + 2\sum_{i=d}^n \varphi_i \Delta\varphi_i - 2\sum_{i=d}^{n-1} \varepsilon_{i+1}\Delta\varphi_i - 2\varepsilon_*\varphi_n - 2\varepsilon_*\Delta\varphi_n}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}\right]$$

We analyze each term of the last equation, neglecting all the terms in order higher than the second in $1/n$ (there is a multiplication in front of z that will make the correction in order $1/n$).

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

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

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

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

$$\mathrm{E}\left[\frac{-2\varepsilon_*\varphi_n}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right] \simeq 0$$

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

We obtain the final result:

$$z = \left[\frac{d(n-d-2)}{(n-d)^2} + \frac{\mu^2}{m^2 n} + \frac{\mu^2}{m^2}\frac{(n-d+1)}{(n+1)^2} + \frac{d}{(n-d+1)} - 2\frac{1}{(n+1)}\frac{\mu}{m}\right]$$

That can be further approximated as:

$$z \simeq \left[2\frac{d}{(n-d)} + 2\frac{\mu^2}{m^2}\frac{1}{n} - 2\frac{1}{n}\frac{\mu}{m}\right]$$