IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

MODEL ORDER SELECTION FROM NOISY POLYNOMIAL DATA WITHOUT USING ANY POLYNOMIAL COEFFICIENTS

Asoke K. Nandi, Fellow, IEEE

Brunel University London, Uxbridge, UB8 3PH, UK
Corresponding author: Asoke K. Nandi (e-mail: asoke.nandi@ brunel.ac.uk)

ABSTRACT Given a set of noisy data values from a polynomial, determining the degree and coefficients of the polynomial is a problem of polynomial regressions. Polynomial regressions are very common in engineering, science, and other disciplines, and it is at the heart of data science. Linear regressions and the least squares method have been around for two hundred years. Existing techniques select a model, which includes both the degree and coefficients of a polynomial, from a set of candidate models which have already been fitted to the data. The philosophy behind the proposed method is fundamentally different to what have been practised in the last two hundred years. In the first stage only the degree of a polynomial to represent the noisy data is selected without any knowledge or reference to its coefficient values. Having selected the degree, polynomial coefficients are estimated in the second stage. The development of the first stage has been inspired by the very recent results that all polynomials of degree q give rise to the same set of known time-series coefficients of autoregressive models and a constant term μ. Computer experiments have been carried out with simulated noisy data from polynomials using four well known model selection criteria as well as the proposed method (PTS1). The results obtained from the proposed method for degree selection and predictions are significantly better than those from the existing methods. Also, it is experimentally observed that the root-mean square (RMS) prediction errors and the variation of the RMS prediction errors from the proposed method scale linearly with the standard deviations of the noise for each degree of a polynomial.

INDEX TERMS data, modelling, model order, polynomial, regression, time-series

I. INTRODUCTION

Polynomial regression aims to select a polynomial that passes near a collection of noisy data values from a polynomial. Polynomial regressions are very common in engineering, science, and other disciplines, and it is one of the important problems of data science. Polynomial regression models are generally fitted with the Least-Squares method to obtain estimated values of the polynomial coefficients. In 1805 Legendre published the Least-Squares method [1] and Gauss published it in 1809 [2] and later in 1823 [35]. In 1815 Gergonne wrote a paper on “The application of the method of least squares to the interpolation of sequences” [3]. This is an English translation by St. John and Stigler [4] of the original paper that was written in French. In the last 120 or so years, polynomial regressions contributed greatly to the development of regression analysis [5-7]. Few more recent and interesting diverse applications can be found in computer graphics [8], machine learning [9], and statistics [10], including robust regressions [11] without the use of the Least-Squares method.

In the last fifty years many model order selection techniques have been developed; the corresponding literature is quite considerable, for example, see [12-13] and references therein. Some of these model order selection techniques are associated with specific model parameter estimation methods and naturally their general applicability is limited. In this paper comparison of the proposed method will be carried out with four model order selection techniques that have been developed around the maximum likelihood method, namely Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), Generalised Information Criterion (GIC), and Bayesian Information Criterion (BIC).

In this paper the focus is on polynomial regressions. Existing techniques select a model, which includes both the degree and all coefficients of a polynomial, from a set of candidate models which have already been fitted to the data. The philosophy behind the proposed method is fundamentally different to what have been done in the last two hundred years. The proposed method for model selection is a two-stage process. In the first stage only the degree of a polynomial to represent the noisy data will be selected without any knowledge or reference to its coefficient values. Having selected the degree, polynomial coefficients will be estimated in the second stage. The first stage has been inspired by the very recent results that all polynomials of degree q give rise to the same set of q known time-series coefficients of autoregressive models and an additional constant term μ [34]. Computer experiments are carried out with simulated noisy data from polynomials using four well known model selection criteria as well as the proposed method to evaluate their accuracies in polynomial degree selection and predictions.

This study is in the context of real-valued and uniformly sampled noisy data from polynomials. The paper presents the following original results:

1) A new and fundamentally different approach to selection of a degree of a polynomial from noisy data, without using any knowledge or reference to the polynomial coefficients, is presented. This is illustrated in section III.
2) New results with noisy data generated from polynomials of degrees 1, 2, 3, and 4 are presented. This can be found in section IV.
3) It is experimentally observed that the model order selection accuracy of the proposed method is the best amongst all the methods compared.
4) Comparison of new results from the existing four methods and the proposed method, are recorded in section IV.
5) The prediction results obtained from the proposed method, PTS1, are significantly better than those from the existing methods. For both the normalised RMS prediction errors and the normalised variations of RMS prediction errors from the proposed method, their relationships with the values of noise standard deviation is linear for each of linear, quadratic, cubic, and quartic polynomials, with a slope of 1.
6) As an internal check of the proposed method, two hidden parameters, specifically $< \mu >$ and its root-mean-square (RMS) error, $\sigma_{\mu(q)}$, have been estimated from the data for each value of the degree of a polynomial and each value of the standard deviation of noise. The agreements between estimated and theoretical values of $< \mu >$ as well as between the estimated and theoretical values of $\sigma_{\mu(q)}$ are found to be remarkable.

II Existing Methods

A polynomial of degree q in continuous time can be written as

$$y(t) = \sum_{i=0}^{q} c(i)\, t^{i} \tag{1}$$

For uniformly sampled discrete time, the continuous time, *t,* is represented as $t = nT$, where n is an integer and T is the sampling period. In this scenario, the above equation can be rewritten as

$$y(nT) = \sum_{i=0}^{q} c(i)\, (nT)^{i}$$

A set of real-valued noisy data from polynomials in uniformly sampled discrete time, can be represented by

$$x(n) = \sum_{i=0}^{q} c(i)\, (n)^{i} + e(n), \tag{2}$$

where $e(n)$ represents errors and T has been removed for the sake of simplicity in notations without the loss of any generalisations as the new value of $c(i)$ is the old value of $c(i)$ multiplied by T^{i}.

In fitting polynomials to such data, $x(n)$, the challenge is to estimate the polynomial coefficients, $c(0), c(1), \dots, c(q)$, as well as the degree of the polynomial, q. One may take the following approach:

1) Choose a value of q.
2) Estimate the corresponding polynomial coefficients, $c(0), c(1), \dots, c(q)$.
3) Calculate the relevant RMS fitting error, $fe(q)$, which is defined as

$$fe(q) = \sqrt{\left[\frac{\left(\sum_{i=1}^{N} \left(\hat{x}(i) - x(i)\right)^{2}\right)}{N}\right]} \tag{3}$$

where N is the number of data values being fitted and $\hat{x}(i)$ are the fitted values.

4) Choose another value of q and go to the step 2 or stop.
5) Find the value of q for which $fe(q)$ is the smallest.
6) Choose this value of q as the estimated degree of the polynomial and the corresponding $c(0), c(1), \dots, c(q)$ as the estimated coefficients of the polynomial.

Unfortunately, this choice is flawed. In general, a larger value of q will result in a lower value of $fe(q)$, and choosing $q = (N-1)$ will always produce a fitting error of $fe(N-1) = 0$, which is often described as overfitting. Therefore, a fitting error, in itself, is not a good indicator of the degree of the underlying polynomial. The aim of model selection techniques is to find a balance between overfitting (too many parameters and zero error) and underfitting (too few parameters and higher error).

There are many model order selection techniques. Four commonly used and well-regarded ones are briefly described below and compared for polynomial data fitting in Section IV and predictions. These are Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), Generalised Information Criterion (GIC), and Bayesian Information Criterion (BIC) [33].

A. Akaike Information Criterion (AIC)

Given a set of models, AIC [13-22] aims to select the best model from this set. Thus, the selected model is not guaranteed to be the best model as it represents a relative choice amongst the given models in the set. AIC tries to balance between the risk of overfitting and the risk of underfitting, as it is a compromise between the best fitted model and the simplicity of the model.

AIC uses the log-likelihood as a measure of the goodness of fit. Suppose that there is a statistical model of the data, that q is the number of estimated parameters, and that L is the maximum value of the likelihood function for the model. AIC is defined as

$$AIC(q) = 2q - 2\ln(L) \tag{4}$$

The first term in equation (4) attempts to keep the polynomial degree small while the second term attempts to obtain the maximum value of the likelihood or the minimum value of the log-likelihood. The selected model will correspond to the one for which AIC is the minimum.

The log-likelihood function for N independent and identical Gaussian distributions is given by

$$lnL\big(\sigma, y(i)\big) = -\left(\frac{N}{2}\right)\ln(\pi) - \left(\frac{N}{2}\right)\ln(\sigma^2) - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{N}\big(x(i) - y(i)\big)^2 \tag{5}$$

For a polynomial of degree 1, i.e., $q = 1$, $y(i) = c(0) + c(1)i$. Thus, $\partial L/\partial c(0) = \sum_{i=1}^{N}(x(i) - c(1)i - c(0))/\sigma^2$,

$\partial L/\partial c(1) = \sum_{i=1}^{N}(x(i) - c(1)i - c(0))i/\sigma^2$, and $\partial L/\partial(\sigma^2) = \left(\frac{N}{2}\right)1/\sigma^2 - \left(\frac{1}{2}\right)1/(\sigma^4)\sum_{i=1}^{N}\big(x(i) - y(i)\big)^2$. The minimum of this log-likelihood function corresponds to the following three equations:

$$\sum_{i=1}^{N}\big(x(i) - c(1)i - c(0)\big) = 0 \tag{6}$$

$$\sum_{i=1}^{N}\big(x(i) - c(1)i - c(0)\big)i = 0 \tag{7}$$

$$\sum_{i=1}^{N}\big(x(i) - y(i)\big)^2 = N\sigma^2 \tag{8}$$

Using equations (6) and (7), one obtains the ordinary least squares estimates of $c(0)$ and $c(1)$. If one sets up two matrices, X^T = [$x(1)$ $x(2)$ ... $x(N)$] and A^T = [1 2 … N; 1 1 … 1], then [$c(1)$ $c(0)$]$^\mathrm{T}$ = $(A^TA)^{-1}A^TX$. It should be remarked that one cannot calculate a value for σ^2 from equation (8) in the absence of the knowledge of the noise-free data, $y(i)$.

Similarly, for a polynomial of degree 2, i.e., $q = 2$, $y(i) = c(0) + c(1)i + c(2)i^2$. Thus, $\partial L/\partial c(0) = \sum_{i=1}^{N}(x(i) - c(2)i^2 - c(1)i - c(0))/\sigma^2$, $\partial L/\partial c(1) = \sum_{i=1}^{N}(x(i) - c(2)i^2 - c(1)i - c(0))i/\sigma^2$, $\partial L/\partial c(2) = \sum_{i=1}^{N}(x(i) - c(2)i^2 - c(1)i - c(0))i^2/\sigma^2$, and $\partial L/\partial(\sigma^2) = \left(\frac{N}{2}\right)1/\sigma^2 - \left(\frac{1}{2}\right)1/(\sigma^4)\sum_{i=1}^{N}\big(x(i) - y(i)\big)^2$. The minimum of this log-likelihood function corresponds to the following four equations:

$$\sum_{i=1}^{N}\big(x(i) - c(2)i^2 - c(1)i - c(0)\big) = 0 \tag{9}$$

$$\sum_{i=1}^{N}\big(x(i) - c(2)i^2 - c(1)i - c(0)\big)i = 0 \tag{10}$$

$$\sum_{i=1}^{N}\big(x(i) - c(2)i^2 - c(1)i - c(0)\big)i^2 = 0 \tag{11}$$

$$\sum_{i=1}^{N}\big(x(i) - y(i)\big)^2 = N\sigma^2 \tag{12}$$

Using equations (9), (10) and (11), one obtains the ordinary least squares estimates of $c(0)$, $c(1)$, and $c(2)$. If one sets up two matrices, X^T = [$x(1)$ $x(2)$ ... $x(N)$] and B^T = [1 4 … N^2; 1 2 … N; 1 1 … 1], then [$c(2)$ $c(1)$ $c(0)$]$^\mathrm{T}$ = $(B^TB)^{-1}B^TX$. Again, it should be remarked that one cannot calculate a value for σ^2 from equation (12) in the absence of the knowledge of the noise-free data, $y(i)$.

Thus, for each selected value of the degree of a polynomial (q), one can estimate the corresponding coefficients of the polynomial, [$c(0)$ $c(1)$ ... $c(q)$], using the above procedure.

B. Corrected Akaike Information Criterion (AICc)

Asymptotically, AIC has certain desirable properties, i.e., as the number of data values tends to ∞. However, whenever the number of data values (N) is small, there is a significant chance that AIC will choose models with too many parameters. This implies that AIC will overfit, despite having the two terms in equation (4), which are intended to offer a balance between underfitting and overfitting. It is worth noting that equation (4) does not have any dependence on the number of data values.

AICc was introduced to address potential overfittings by Sugiura [23] in the context of linear regression. Since then Hurvich and Tsai [24] as well as many other researchers, e.g., [13], [16], [20], and [25], have extended the applicability of AICc. AICc can be defined as

$$AICc(q) = \frac{2qN}{N - q - 1} - 2\ln(L) \tag{13}$$

It is clear from equation (13) that AICc depends, amongst other factors, on the number of data values (N). As in AIC, AICc also attempts to find a balance between underfitting and overfitting.

The procedure for selecting the best model from a given set of models, i.e., the degree of a polynomial (q) and the corresponding coefficients of the polynomial, $[c(0)\ c(1) \dots c(q)]$, requires the minimisation of AICc and is essentially the same as outlined in section IIA. Similar observations, as for AIC, can be made for AICc. While, for each value of the degree of a polynomial (q), one can estimate the corresponding coefficients of the polynomial, $[c(0)\ c(1) \dots c(q)]$, using the above procedure, it should be remarked that one cannot calculate a value for σ^2 in the absence of the knowledge of the noise-free data, $y(i)$.

C. Generalised Information Criterion (GIC)

In AIC, the factor of $2q$ has been designed to address the issue of overfitting. Intuitively, the probability of overfitting will be reduced as the number of data values increases. In finite sample situations, extensive simulation studies have demonstrated that the following generalised information criterion (GIC) [26]

$$GIC(q) = \alpha q - 2\ln(L) \tag{14}$$

can perform better than AIC if $\alpha > 2$. Values of α in the range from 2 to 6 appear to offer the best performance. The optimal value for α appears to depend on many factors and there is no clear hint on how to choose its value in a specific scenario. Note that $\alpha = 2$ corresponds to AIC. In the following investigations the value for α has been chosen to be 4. Note that GIC does not explicitly depend on the number of data values.

The procedure for selecting the best model from a given set of models, i.e., the degree of a polynomial (q) and the corresponding coefficients of the polynomial, $[c(0)\ c(1) \dots c(q)]$, requires the minimisation of GIC and is essentially the same as outlined in section IIA. Again, similar observations, as for AIC and AICc, can be made for GIC. While, for each value of the degree of a polynomial (q), one can estimate the corresponding coefficients of the polynomial, $[c(0)\ c(1) \dots c(q)]$, using the above procedure, it should be remarked that one cannot calculate a value for σ^2 in the absence of the knowledge of the noise-free data, $y(i)$.

D. Bayesian Information Criterion (BIC)

The form of BIC [27-32], [13], [16] is very similar to AIC, in that they both have two terms – a negative log-likelihood one and a penalty term for the number of parameters. However, their origins are different. The log-likelihood term is identical in both cases. The penalty term is $2q$ in AIC, while it is $\ln(N)\,(q)$ in BIC. Note that AIC does not depend on the number of data values (N), but BIC does include a dependence on N. In that sense, BIC captures something of AIC and AICc. BIC can be written as

$$BIC(q) = \ln(N)\,(q) - 2\ln(L) \tag{15}$$

The procedure for selecting the best model from a given set of models, i.e., the degree of a polynomial (q) and the corresponding coefficients of the polynomial, $[c(0)\ c(1) \dots c(q)]$, requires the minimisation of BIC and is essentially the same as outlined in section IIA. Similar observations, as for AIC, AICc, and GIC, can be made for BIC. While, for each value of the degree of a polynomial (q), one can estimate the corresponding coefficients of the polynomial, $[c(0)\ c(1) \dots c(q)]$, using the above procedure, it should be remarked that one cannot calculate a value for σ^2 in the absence of the knowledge of the noise-free data, $y(i)$.

Yang [31] compared AIC and BIC in the context of regression under the assumption that the true model is not present in the set of models being compared. It was demonstrated that AIC was asymptotically optimal for selecting the model with the least mean squared error, while BIC was not asymptotically optimal under the same assumption. In the following investigations, the true model order is present in the set of models being compared though not the exact coefficient values but only the estimated values from ordinary Least-Squares (OLS).

When the true model is present in the set of models being compared, it is well documented that BIC will select the true model with probability 1 asymptotically, i.e., as N tends to ∞. Vrieze [32] carried out a simulation study including the true model in the set of models being compared. It was shown that AIC can sometimes choose a much better model than BIC, since there is a significant chance of BIC choosing a bad model for finite values of N.

III Proposed Method

The proposed method is fundamentally different from all existing methods. All existing methods, including AIC, AICc, GIC, and BIC, selects one of the given polynomial models from a given a set of models where each model consists of a value for the degree of a polynomial as well as values for its coefficients. It should be noted that the knowledge of coefficient values of the polynomial are needed to generate the log-likelihood estimates [33]. Therefore, these methods require both the values of the estimated polynomial coefficients, $c(0), c(1), \dots, c(q)$, as well as the degree of the polynomial, q.

It is important to note that in the proposed method there are two stages. In the first stage, the selection of only the degree of the polynomial is carried out without the knowledge of the values of the corresponding coefficients. Having selected the degree of the polynomial in the first stage, the polynomial coefficients are calculated in the second stage. The first stage is certainly novel and may seem impossible, as there are infinitely many polynomials of any one value for the degree. Nonetheless, this has been inspired by the very recent results

[34], as shown below, which proved that all data from uniformly sampled polynomials of finite degree q can be perfectly represented by an autoregressive time-series model of order q such that

$$y(n) = \sum_{i=1}^{q} a(i)\, y(n-i) + \mu \qquad (16)$$

where

$$a(i) = (-1)^{i+1} \binom{q}{i} \qquad (17)$$

for $i = 1, 2, \dots, q$, and

$$\mu = c(q)(q!) \qquad (18)$$

A. Selection of the degree of a polynomial

Equation (16) can be rewritten as

$$-\sum_{i=1}^{q} a(i)\, y(n-i) + y(n) = \mu$$

As noise-free $y(n)$ values are not available, this equation is recast with known noisy data values $x(n)$ as follows

$$-\sum_{i=1}^{q} a(i)\, x(n-i) + x(n) = \mu(n, q) \qquad (19)$$

where $\mu(n, q)$ may depend on both n and q. Equation (19) can be written in matrix form as XA = M, where X is a $(f - q)$ x $(q + 1)$ matrix and X = $[\ \big(x(1) \dots x(f-q)\big)^T\ ;\ \big(x(2) \dots x(f-q+1)\big)^T;\ \dots\ \big(x(q+1) \dots x(f)\big)^T]$, A is a $(q+1)$x1 matrix and A = $[(-a(1) \dots - a(q)\ 1)^T]$, M is a $(f-q)$x1 matrix and M = $[\mu(q+1, q) \dots \mu(f, q))^T]$, as well as f is the number of data values being used for estimation.

All the entries in matrix X are known as they represent the noisy data values. Also, all the entries in matrix A are known from equation (17). Therefore, M can be obtained from XA, containing $(f - q)$ values. For a chosen value of q, these $(f - q)$ values are estimates of the constant term, $\mu(q)$, for a polynomial of degree q. From these $(f - q)$ values, one can estimate the mean value, $<\mu(q)>$, and the root-mean square value, $\sigma_{\mu(q)}$, of the $\mu(q)$.

Now, equation (19) can be rearranged and approximated as follows

$$\hat{x}(n) = \sum_{i=1}^{q} a(i)\, x(n-i) + <\mu(q)>$$

As everything on the right hand of the above equation is known, these $(f - q)$ values of $\hat{x}(n)$ are calculated and can be regarded as time-series "fitted" values. The relevant root-mean square time-series estimation error, $fe(q)$, is defined as

$$fe(q) = \sqrt{\left[\frac{\left(\sum_{i=q}^{f} \big(\hat{x}(i) - x(i)\big)^2\right)}{(f-q)}\right]} \qquad (20)$$

where $(f - q)$ is the number of data values being estimated and $\hat{x}(i)$ are the estimated values. It should be noted that $fe(q)$ generally decreases as q increases.

Thus, for every value of q, there are three parameters - the mean value, $<\mu(q)>$, the root-mean square value, $\sigma_{\mu(q)}$, of the $\mu(q)$, and the root-mean square time-series estimation error, $fe(q)$. These are used to select the appropriate value of q for the noisy polynomial data.

Recall that each of AIC, AICc, GIC, and BIC attempts to balance between overfitting and underfitting scenarios. In one scenario errors reduce and in the other scenario errors increase with increasing values of q. A similar scenario arises here, in that $fe(q)$ generally decreases as q increases, while $\sigma^2_{\mu(q)}$ increases with q. Now the following parameter is defined

$$PTS1(q) = \big(fe(q)\big)^2 + \sigma^2_{\mu(q)} \qquad (21)$$

The selected value of q is the one for which PTS1(q) is the minimum. This is the proposed selection criterion.

B. Estimation of polynomial coefficients

In section IIIA, the appropriate degree, q, for the noisy polynomial data has already been selected. Now equation (2) can be written in matrix form as GC = X, where G is a fx$(q+1)$ matrix and G = $[(1\ 2^q \dots f^q)^T;\ \dots\ (1\ 2\ \dots f)^T; (1 \dots 1)^T]$, C is a $(q+1)$x1 matrix and C = $[(c(q) \dots c(0))^T]$, X is a fx1 matrix and X = $[x(1)\ \dots\ x(f))^T]$, as well as f is the number of data values being used for estimation.

All the entries in matrix G are known as they represent integer powers of integers (see equation (2)). Also, all the entries in matrix X are known as they represent noisy data values (see equation (2)). Therefore, C, containing the coefficients of the polynomial of the selected degree, can be estimated. As the matrix G is not square, one can obtain the Ordinary Least Squares (OLS) solution using the pseudoinverse of G, namely,

$$C = inv(G^T G) G^T X \qquad (22)$$

Thus, for the selected value of q, the matrix C contains estimated values of the q coefficients of the polynomial.

C. Predictions

In section III A and B, the appropriate degree, q, for the noisy polynomial data has already been selected and the corresponding polynomial coefficients have been estimated using the first f data values. Using this information, the remaining $(N-f)$ data values are predicted. The predicted data values are obtained using a similar matrix equation to $GC = \hat{X}$, where G is a $(N-f)\text{x}(q+1)$ matrix and $G = [((f+1)^q \dots N^q)^T;\ \dots\ ((f+1)\ \dots N)^T; (1 \dots 1)^T]$, C is a $(q+1)\text{x}1$ matrix and C = $[(c(q) \dots c(0))^T]$, X is a $f\text{x}1$ matrix and $\hat{X} = [\hat{x}(f+1) \dots\ \hat{x}(N))^T]$, as well as $(N-f)$ is the number of data values being used for prediction. The relevant root-mean square prediction errors, $pe(q)$, defined as

$$pe(q) = \sqrt{\left[\frac{\left(\sum_{i=f+1}^{N}\left(\hat{x}(i) - x(i)\right)^2\right)}{(N-f)}\right]} \quad (23)$$

where $(N-f)$ is the number of data values being predicted and $\hat{x}(i)$ are the fitted values, have been calculated.

IV Results

In this section are described some computer experiments to assess the performance of the four existing techniques (AIC, AICc, GIC, and BIC) as well as the proposed method (PTS1) for selecting polynomial models and predicting noisy polynomial data. Linear, quadratic, cubic, and quartic polynomials with different amounts of Gaussian noise have been considered. It is clear from equations (4), (13), (14), and (15) that AIC, AICc, GIC, and BIC calculate log-likelihoods which require the knowledge of the standard deviation of noise, which is not available in real situations. One can attempt to get a reasonable estimate of this using the fitted values from the model and the noise-free data using equation (8). Since noise-free data are not available in real situations, one can attempt to estimate the standard deviations of noise.

Earlier experiments using estimated standard deviations have produced poor results from AIC, AICc, GIC, and BIC. Therefore, in the following experiments, AIC, AICc, GIC, and BIC results are based on using the exact values of the standard deviations of noise (known from simulations), which give them an advantage over the proposed method, PTS1, that does not use such exact knowledge.

A. Linear polynomial

Here a polynomial of degree 1 has been considered for data generation: $y(n) = n + 1$. For this experiment, 1,000 sets of 101 data values have been generated for each value of the standard deviation (σ) of noise using the zero-mean Gaussian distribution, $\mathbb{N}(0,\sigma)$. Thus, the generated noisy data can be described by

$$x(n,\sigma) = n + 1 + \mathbb{N}(0,\sigma),$$

$$\text{for } n = -50{:}\,1{:}\,50 \text{ and } \sigma = 1{:}\,1{:}\,5 \quad (24)$$

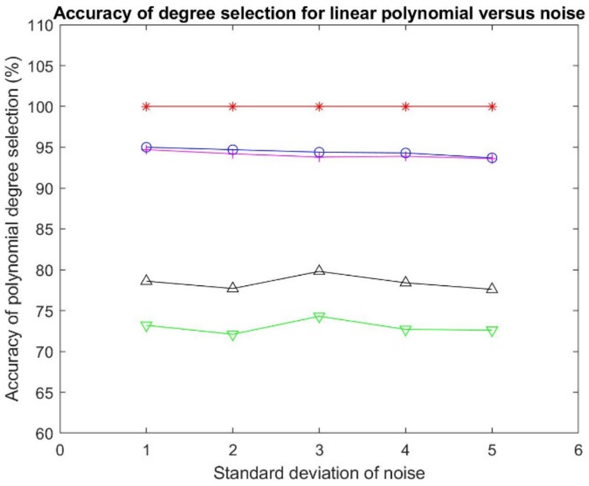

Figure 1. The accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for a linear polynomial.

In each set of 101 data values, the first 60 data values have been used for estimating the degree of the polynomial and its coefficients. Figure 1 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles),

Table 1. Root mean-square errors for predicting $x(n) = n + 1 + \mathbb{N}(0,\sigma)$ using five different techniques, including the proposed technique

Degree of polynomial, q	Standard deviation of noise, σ	AIC (exact σ)	AICc (exact σ)	GIC (exact σ)	BIC (exact σ)	Proposed technique (estimated σ)
1	1	124	93.9	11.7	11.7	1.08
1	2	241	131	5.77	4.11	2.17
1	3	339	238	43.1	43.1	3.24
1	4	561	420	15.1	12.3	4.35
1	5	543	406	24.7	23.2	5.41

and PTS1 (red stars) versus noise standard deviations for the linear polynomial. AICc is always better than AIC. GIC and BIC are very similar, and they are always much better than AIC and AICc. The proposed method, PTS1, is always the best by far.

The remaining 41 data values have been predicted and the root-mean square (RMS) prediction errors have been calculated. The RMS prediction errors for AIC, AICc, GIC, BIC, and the proposed method for each of the five values of the standard deviation of noise are presented in Table 1.

Of the existing techniques AIC, AICc, GIC, and BIC, GIC and BIC offer much better performance. Each of AIC, AICc, GIC, and BIC calculates log-likelihoods, which require, amongst others, the value of σ, and was allowed the advantage of using the exact values of σ for the above results, even though these are not available in reality. Despite this, the proposed technique is the best and offers significantly better performance than GIC and BIC (as well as the other two).

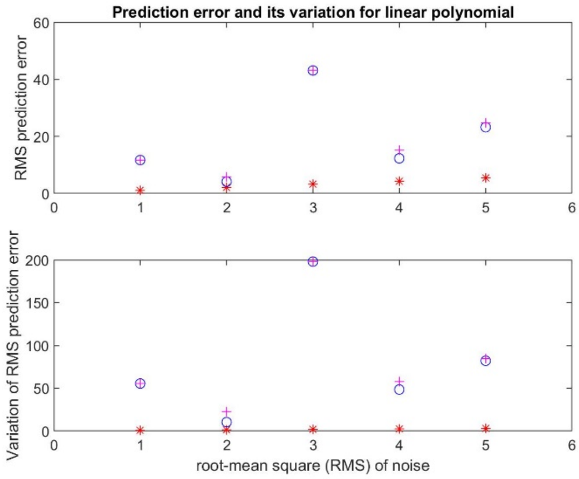

Figure 2. The upper figure shows the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), and the lower figure shows the variation of RMS prediction errors versus the standard deviations of noise for a linear polynomial.

The upper half of Figure 2 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), while the lower half of Figure 2 depicts the variation of RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars). It is remarkable how much smaller the RMS prediction errors from the proposed method are compared to those from GIC and BIC. Also, the variations of RMS prediction errors from the proposed method are significantly smaller compared to those from GIC and BIC.

B. Quadratic polynomial

Here a polynomial of degree 2 has been considered for data generation: $y(n) = n^2 + n + 1$. For this experiment, 1,000 sets of 101 data values have been generated for each value of the standard deviation (σ) of noise using the zero-mean Gaussian distribution, $\mathbb{N}(0, \sigma)$. Thus, the generated noisy data can be described by

$$x(n, \sigma) = n^2 + n + 1 + \mathbb{N}(0, \sigma),$$

$$\text{for } n = -50{:}\,1{:}\,50 \text{ and } \sigma = 1{:}\,1{:}\,5, \qquad (25)$$

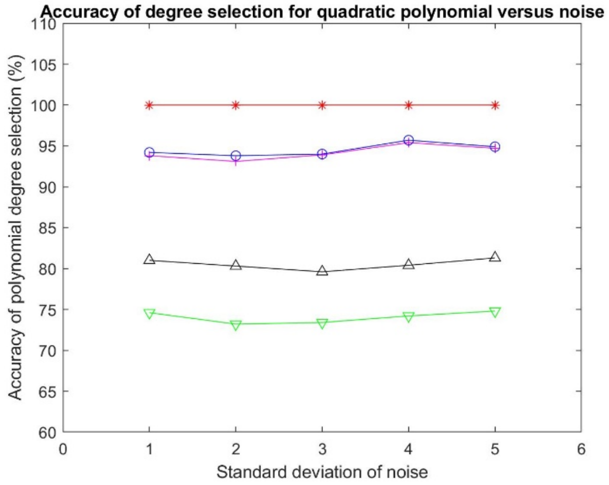

Figure 3. The accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for a quadratic polynomial.

In each set of 101 data values, the first 60 data values have been used for estimating the degree of the polynomial and its coefficients. Figure 3 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black

Table 2. Root mean-square errors for predicting $x(n) = n^2 + n + 1 + \mathbb{N}(0, \sigma)$ using five different techniques, including the proposed technique.

Degree of polynomial, q	Standard deviation of noise, σ	AIC (exact σ)	AICc (exact σ)	GIC (exact σ)	BIC (exact σ)	Proposed technique (estimated σ)
2	1	127	90.9	16.5	16.4	1.66
2	2	273	191	30.0	26.5	3.32
2	3	425	332	162	160	5.16
2	4	593	470	132	129	6.70
2	5	729	532	220	220	8.49

upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for the quadratic polynomial. AICc is always better than AIC. GIC and BIC are very similar, and they are always much better than AIC and AICc. The proposed method, PTS1, is always the best by far.

The remaining 41 data values have been predicted and the root-mean square (RMS) prediction errors have been calculated. The RMS prediction errors for AIC, AICc, GIC, BIC, and the proposed method for each of the five values of the standard deviation of noise are presented in Table 2.

Of the existing techniques AIC, AICc, GIC, and BIC, GIC and BIC offer much better performance. Each of AIC, AICc, GIC, and BIC calculates log-likelihoods, which require, amongst others, the value of σ, and was allowed the advantage of using the exact values of σ for the above results, even though these are not available in reality. Despite this, the proposed technique is the best and offers significantly better performance than GIC and BIC, typically an order of magnitude better.

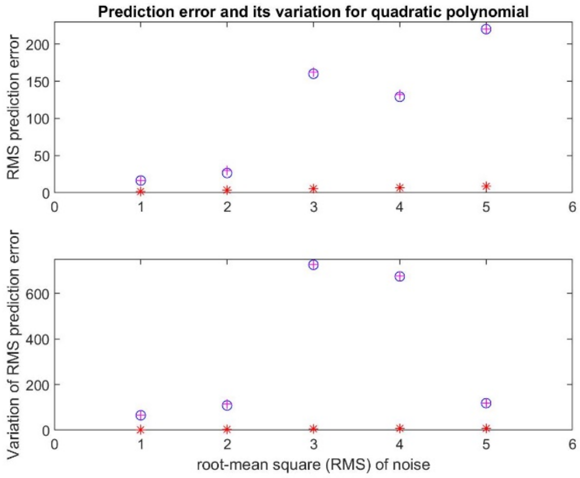

Figure 4. The upper figure shows the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), and the lower figure shows the variation of RMS prediction errors versus the standard deviations of noise for a quadratic polynomial.

The upper half of Figure 4 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles) and the proposed method (red stars), while the lower half of Figure 4 depicts the variation of RMS prediction errors from GIC (magenta + signs), BIC (blue circles) and the proposed method (red stars). It is remarkable how much smaller the RMS prediction errors from the proposed method are compared to those from GIC and BIC. Also, the variations of RMS prediction errors from the proposed method are significantly smaller compared to those from GIC and BIC.

C. Cubic polynomial

Here a polynomial of degree 3 has been considered for data generation: $y(n) = n^3 + n^2 + n + 1$. For this experiment, 1,000 sets of 101 data values have been generated for each value of the standard deviation (σ) of noise using the zero-mean Gaussian distribution, $\mathbb{N}(0, \sigma)$. Thus, the generated noisy data can be described by

$$x(n, \sigma) = n^3 + n^2 + n + 1 + \mathbb{N}(0, \sigma),$$
$$\text{for } n = -50{:}\,1{:}\,50 \text{ and } \sigma = 1{:}\,1{:}\,5, \tag{26}$$

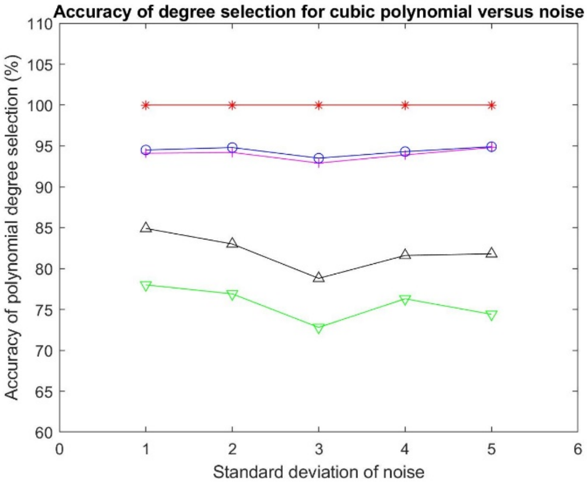

Figure 5. The accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations a cubic polynomial.

In each set of 101 data values, the first 60 data values have been used for estimating the degree of the polynomial and its coefficients. Figure 5 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for the cubic polynomial. AICc is always better than AIC. GIC and BIC are very similar, and they are always much better than AIC and AICc. The proposed method, PTS1, is always the best by far.

The remaining 41 data values have been predicted and the root-mean square (RMS) prediction errors have been calculated. The RMS prediction errors for AIC, AICc, GIC, BIC, and the proposed method for each of the five values of the standard deviation of noise are presented in Table 3.

Of the existing techniques AIC, AICc, GIC, and BIC, GIC and BIC offer much better performance. Each of AIC, AICc, GIC, and BIC calculates log-likelihoods, which require, amongst others, the value of σ, and was allowed the advantage of using the exact values of σ for the above results, even though these are not available in reality. Despite this, the proposed technique is the best and offers significantly better performance than GIC and BIC (as well as the other two).

Table 3. Root mean-square errors for predicting $x(n) = n^3 + n^2 + n + 1 + \mathbb{N}(0, \sigma)$ using five different techniques, including the proposed technique.

Degree of polynomial, N	Standard deviation of noise, σ	AIC (exact σ)	AICc (exact σ)	GIC (exact σ)	BIC (exact σ)	Proposed technique (estimated σ)
3	1	140	90.5	41.9	30.8	5.19
3	2	323	264	122	122	10.2
3	3	565	489	252	251	15.1
3	4	561	452	193	183	20.1
3	5	904	630	215	199	25.6

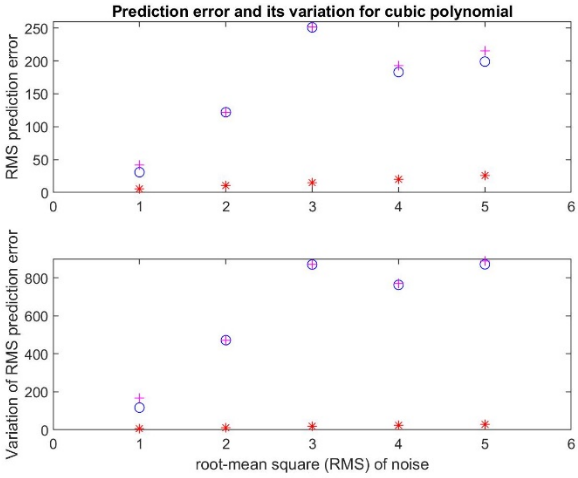

Figure 6. The upper figure shows the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), and the lower figure shows the variation of RMS prediction errors versus the standard deviations of noise for a cubic polynomial.

The upper half of Figure 6 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles) and the proposed method (red stars), while the lower half of Figure 6 depicts the variation of RMS prediction errors from GIC (magenta + signs), BIC (blue circles) and the proposed method (red stars). It is remarkable how much smaller the RMS prediction errors from the proposed method are compared to those from GIC and BIC. Also, the variations of RMS prediction errors from the proposed method are significantly smaller compared to those from GIC and BIC.

D. Quartic polynomial

Here a polynomial of degree 4 has been considered for data generation: $y(n) = n^4 + n^3 + n^2 + n + 1$. For this experiment, 1,000 sets of 101 data values have been generated for each value of the standard deviation (σ) of noise using the zero-mean Gaussian distribution, $\mathbb{N}(0, \sigma)$. Thus, the generated noisy data can be described by

$$x(n, \sigma) = n^4 + n^3 + n^2 + n + 1 + \mathbb{N}(0, \sigma),$$

$$\text{for } n = -50{:}\,1{:}\,50 \text{ and } \sigma = 1{:}\,1{:}\,5, \qquad (27)$$

In each set of 101 data values, the first 60 data values have been used for estimating the degree of the polynomial and its coefficients. Figure 7 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for the quartic polynomial. AICc is always better than AIC. GIC and BIC are very similar, and they are always much better than AIC and AICc. The proposed method, PTS1, is always the best by far.

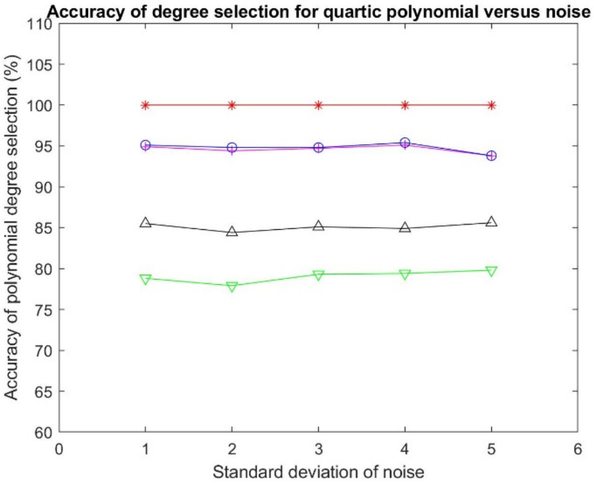

Figure 7. The accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for a quartic polynomial.

The remaining 41 data values have been predicted and the root-mean square (RMS) prediction errors have been calculated. The RMS prediction errors for AIC, AICc, GIC, BIC, and the proposed method for each of the five values of the standard deviation of noise are presented in Table 4.

Of the existing techniques AIC, AICc, GIC, and BIC, GIC and BIC offer much better performance. Each of AIC, AICc, GIC, and BIC calculates log-likelihoods, which require, amongst others, the value of σ, and was allowed the advantage of using the exact values of σ for the above results, even though these are not available in reality. Despite this, the proposed technique is the best and offers significantly better performance than GIC and BIC (as well as the other two).

Table 4. Root mean-square errors for predicting $x(n) = n^4 + n^3 + n^2 + n + 1 + \mathbb{N}(0, \sigma)$ using five different techniques, including the proposed technique.

Degree of polynomial, *N*	Standard deviation of noise, σ	AIC (exact σ)	AICc (exact σ)	GIC (exact σ)	BIC (exact σ)	Proposed technique (estimated σ)
4	1	209	179	121	121	19.7
4	2	415	356	214	213	38.7
4	3	658	566	330	329	58.8
4	4	855	731	429	406	81.4
4	5	947	808	582	582	94.6

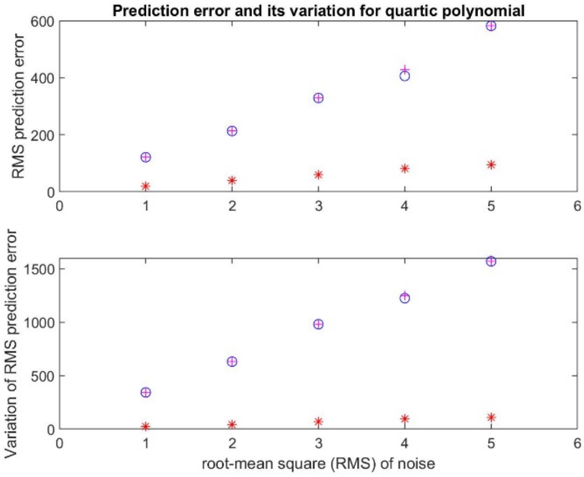

Figure 8. The upper figure shows the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), and the lower figure shows the variation of RMS prediction errors versus the standard deviations of noise for a quartic polynomial.

The upper half of Figure 8 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles) and the proposed method (red stars), while the lower half of Figure 8 depicts the variation of RMS prediction errors from GIC (magenta + signs), BIC (blue circles) and the proposed method (red stars). It is remarkable how much smaller the RMS prediction errors from the proposed method are compared to those from GIC and BIC. Also, the variations of RMS prediction errors from the proposed method are significantly smaller compared to those from GIC and BIC.

E. Normalised results from the proposed method

Figure 9 displays the normalised RMS prediction errors and the normalised variation of RMS prediction errors from the proposed method for linear, quadratic, cubic, and quartic polynomials. All the RMS prediction errors from the proposed method for the linear polynomial were divided by its RMS error at $\sigma = 1$. This ensured that the normalised RMS prediction error for the linear polynomial $\sigma = 1$ is 1. Similarly, all the variations of RMS prediction errors from the proposed method for the linear polynomial were divided by its value at $\sigma = 1$. This ensured that the normalised variation of RMS prediction error for the linear polynomial $\sigma = 1$ is 1. Similar procedures were repeated for quadratic, cubic, and quartic polynomials.

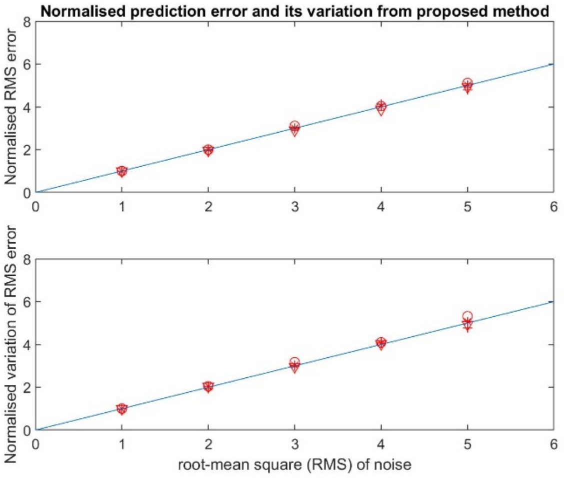

Figure 9. The upper figure shows the normalised RMS prediction errors from the proposed method for linear (red stars), quadratic (red open circles), cubic (red upper triangles), and quartic (red plus signs) polynomials. The lower figure shows the normalised variations of RMS prediction errors versus the standard deviations of noise. The two blue straight lines with the slope of 1 and the intercept of 0 are for guidance.

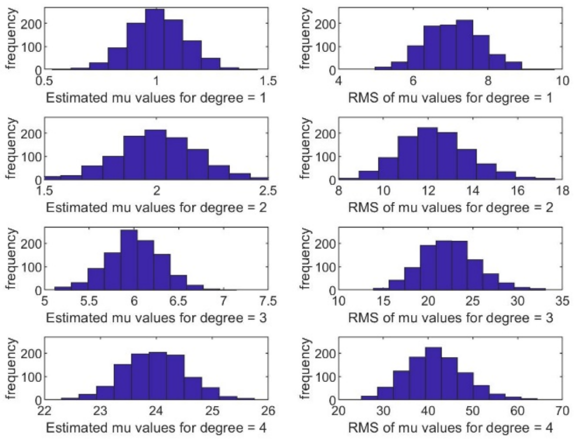

Figure 10. The left column shows the histograms of 1000 values of <μ> for degrees ranging from 1 to 4. The right column shows the histograms of 1000 values of $\sigma_{\mu(q)}$ for degrees ranging from 1 to 4.

Table 5. Estimated and theoretical values of $< \mu >$ and $\sigma_{\mu(q)}$ for different polynomial degrees and noise, σ, of 5.

Degree of polynomial, q	Standard deviation of noise, σ	Estimated $< \mu >$	Theoretical $< \mu >$	Estimated $\sigma_{\mu(q)}$	Theoretical $\sigma_{\mu(q)}$
1	5	1.0	1.0	7.1	7.1
2	5	2.0	2.0	12.3	12.2
3	5	6.0	6.0	22.4	22.4
4	5	24.0	24.0	41.7	41.8

The upper half of Figure 9 displays the normalised RMS prediction errors from the proposed method for linear (marked by red stars), quadratic (marked by red open circles), cubic (marked by red upper triangles), and quartic (marked by red plus signs) polynomials. The description of the lower half of Figure 9 is essentially the same, except that these values represent the normalised variations of RMS prediction errors. The two blue straight lines in Figure 9 have a slope of 1 and an intercept of 0 for guidance.

There are three remarkable observations from Figure 9:

1) For both the normalised RMS prediction errors and the normalised variations of RMS prediction errors from the proposed method, their relationships with the values of σ appear to be linear for each of linear, quadratic, cubic, and quartic polynomials.
2) For both the normalised RMS prediction errors and the normalised variations of RMS prediction errors from the proposed method, their relationships with the values of σ for each of linear, quadratic, cubic, and quartic polynomials appear to be identical.
3) For both the normalised RMS prediction errors and the normalised variations of RMS prediction errors from the proposed method, their relationships with the values of σ may be described by a single line with a slope of 1.

F. μ values and RMS of μ values

In [34], it has been proven that, for the correct degree, $\mu = c(q)\,(q!)$, for noise-free data. Taking the expectation of equation (19), one can write

$$< \mu > = < \mu(q) > = < \langle -\sum_{i=1}^{q} a(i)\, x(n-i) \; + x(n) \rangle >$$

$$= -\sum_{i=1}^{q} a(i)\, y(n-i) \; + y(n) = c(q)\,(q!) \qquad (28)$$

Since in these experiments $c(q) = 1$ in all cases, the theoretical expectations are that $< \mu > = q!$.

Given that $\mu(n,q) = x(n) - \sum_{i=1}^{q} a(i)\, x(n-i)$, variations in $\mu(n,q)$ values will come, in the current scenario of having chosen the correct degree, from variations in $x(n)$. Therefore, the variance is given by

$$< \sigma_{\mu(q)} >^2 = \left((1)^2 + \sum_{i=1}^{q} a(i)^2 \right) \sigma^2 \qquad (29)$$

Combining equations (17) and (29), theoretically expected values of $< \sigma_{\mu(q)} >^2$ are 50, 150, 500, and 1750, for polynomial degree (q) values of 1, 2, 3, and 4 respectively for noise standard deviation of 5.

In computer experiments there are 1,000 realisations for each degree of polynomial and each value of noise standard deviation, σ. In each of these 1,000 realisations there is one value for $< \mu >$ and one value for $\sigma_{\mu(q)}$. Figure 10 has four rows and two columns. All the eight subplots correspond to σ = 5. Each column contains four subplots. The first column displays histograms of 1,000 values of $< \mu >$ for degree = 1 in the top subplot, for degree = 2 in the next subplot, for degree = 3 in the next subplot, and for degree = 4 in the bottom subplot. The second column displays histogram of 1,000 values of $\sigma_{\mu(q)}$ for degree = 1 in the top subplot, for degree = 2 in the next subplot, for degree = 3 in the next subplot, and for degree = 4 in the bottom subplot. All eight distributions look fairly symmetrical.

Table 5 records the average of these 1,000 values and RMS values from computer experiments as well as the corresponding theoretical values for each value of degree and σ = 5. Theoretical values of $< \mu >$ for different values of q can be calculated from equation (28) and theoretical values of $\sigma_{\mu(q)}$ can be calculated from equation (29). From Table 5 it is clear that the agreements between estimated and theoretical values of $< \mu >$ as well as between estimated and theoretical values of $\sigma_{\mu(q)}$ are remarkable.

V Discussion

In the above experiments, for model selections and predictions, AIC results are the worst and AICc results are better than AIC results. GIC and BIC results are very similar, and they are much better than AIC and AICc results. Remarkably, results from the proposed method are the best and significantly better than GIC and BIC results.

One natural question is “what is the statistical significance of the polynomial degree estimation results from the various estimators for the four different polynomial degrees and the five different noise standard deviations”. For each combination of a polynomial degree and a noise standard deviation, there are 1000 estimated values of the polynomial degree for each of the five estimators. As estimated polynomial degrees are discrete, the "poissfit" function in MATLAB has been used to obtain the maximum likelihood estimate of the degree and its 95% confidence interval (i.e., the significance level of 0.05). The maximum likelihood estimate of the degree in all 20 cases for each of the four existing estimators AIC, AICc, GIC, and BIC is different from the true value of the degree. Table 6 contains the results from these four estimators, describing either the true polynomial degree is inside or outside the 95% confidence intervals. For AIC and AICc, confidence intervals do not contain the true degree 100% of these cases. For GIC and BIC, confidence intervals do not contain the true degree 25% of these cases. On the other hand, PTS1 has selected the correct polynomial degree in all of these 20 cases (4 degrees * 5 noise standard deviations). Thus, PTS1 results are consistent with the true degree for 100% of these cases in the presence of zero-mean Gaussian noise.

Table 6. Is the true polynomial degree inside or outside the 95% confidence intervals for different estimators at four different polynomial degrees and five different zero-mean Gaussian noise standard deviations?

Polynomial degree	noise	AIC	AICc	GIC	BIC
1	1	Outside	Outside	Outside	Outside
1	2	Outside	Outside	Outside	Outside
1	3	Outside	Outside	Outside	Outside
1	4	Outside	Outside	Outside	Outside
1	5	Outside	Outside	Outside	Outside
2	1	Outside	Outside	Inside	Inside
2	2	Outside	Outside	Inside	Inside
2	3	Outside	Outside	Inside	Inside
2	4	Outside	Outside	Inside	Inside
2	5	Outside	Outside	Inside	Inside
3	1	Outside	Outside	Inside	Inside
3	2	Outside	Outside	Inside	Inside
3	3	Outside	Outside	Inside	Inside
3	4	Outside	Outside	Inside	Inside
3	5	Outside	Outside	Inside	Inside
4	1	Outside	Outside	Inside	Inside
4	2	Outside	Outside	Inside	Inside
4	3	Outside	Outside	Inside	Inside
4	4	Outside	Outside	Inside	Inside
4	5	Outside	Outside	Inside	Inside

Another question is "how does PTS1 work for a non-Gaussian noise distribution". There are very many non-Gaussian distributions. A new set of 20 experiments were carried out using zero-mean noise from a Uniform distribution at the same five noise standard deviations and four polynomial degrees. For each of the five estimators, there are now 20 (4 degrees * 5 noise standard deviations) pairs of numbers; one number is the percentage accuracy with Gaussian noise while the other number is the percentage accuracy with Uniform noise. For each estimator, 20 values of the percentage accuracy with Gaussian noise minus the corresponding percentage accuracy with Uniform noise have been calculated. For each estimator, both the average of these difference percentage accuracies and the standard deviation of these difference percentage accuracies have been calculated from these 20 values. The results are 0.65% ± 2.10% for AIC, -0.24% ± 1.78% for AICc, -0.37% ± 0.97% for GIC, -0.37% ± 0.97% for BIC, and 0.0% ± 0.0% for PTS1. This confirms that the results with Gaussian noise and Uniform noise are very similar for each estimator. Again, the PTS1 is the only one to select the correct degree of a polynomial every time; its performance was always the best. BIC and GIC performances are very similar, and their performances are better than AICc, which performed better than AIC.

For Uniform noise, the same "poissfit" function in MATLAB has been used to obtain the maximum likelihood estimate of the degree and its 95% confidence interval (i.e., the significance level of 0.05). The maximum likelihood estimate of the degree in all 20 cases for each of the four existing estimators is different from the true value of the degree. For AIC and AICc, confidence intervals do not contain the true polynomial degree 100% of these cases. For GIC and BIC, confidence intervals do not contain the true degree 25% of these cases. On the other hand, PTS1 has selected the correct polynomial degree in all of these 20 cases (4 degrees * 5 noise standard deviations). Thus, PTS1 results are consistent with the true degree for 100% of these cases in the presence of zero-mean Uniform noise.

Recall that each of AIC, AICc, GIC, and BIC attempts to balance between overfitting and underfitting scenarios. In one scenario errors reduce and in the other scenario errors increase with increasing values of q. As can be seen from equations (4), (13), (14), and (15), each of them has the identical log-likelihood term, wherein one can find the fitting error. At a conceptual plane, the $\left(fe(q)\right)^2$ term in PTS1(q) (see equation (21)) is similar in that it also represents "fitting" error (but not the same at all). The first term in these four existing estimators always depends on q and on N for AICc and BIC. It is worth noting that this first term has no connection with the data values. On the contrary, the first term in equation (21) for PTS1(q) is $\sigma^2_{\mu(q)}$, which clearly connected with the underlying data values (beyond q and N). This is yet another way that the proposed estimator is different from other four estimators.

A great deal of attempts has been made to make the comparisons fair (see section II and IV). However, AIC, AICc, GIC, and BIC calculate log-likelihoods which require the knowledge of the standard deviation of noise, which is not available in real situations. One can attempt to get a reasonable estimate of this using the fitted values from the model as the noise-free data using equation (8). Earlier experiments using estimated standard deviations have produced poor results from AIC, AICc, GIC, and BIC. Therefore, in all the above experiments, AIC, AICc, GIC, and BIC results are based on using the exact values of the standard deviations of noise (known from simulations), which give them an advantage over the proposed method, PTS1.

Just as in every study, there are some limitations of this study. It is important to remember that the observations have been based on zero-mean, symmetric noise distributions. It is not known how things will work out in non-symmetric or other (than Gaussian and Uniform) symmetric noise distributions. Also, these experiments have been performed with four different polynomial degrees and five different noise standard deviations. Extensions of these may be considered in future explorations.

VI Conclusion

Given a set of noisy data values from a polynomial, this paper has considered determining the degree and coefficients of the polynomial, which is the problem of polynomial regressions. Unlike existing techniques, which select a model, including both the degree and coefficients of a polynomial, from a set of candidate models which have already been fitted to the data, the proposed method is fundamentally different. In the first stage only the degree of a polynomial to represent the noisy data is selected without any knowledge or reference to its coefficient values. Having selected the degree, polynomial coefficients are estimated in the second stage. The first stage has been inspired by the very recent results that all polynomials of degree q give rise to the same set of known time-series coefficients of autoregressive models and a constant term μ [34] and it constitutes a different paradigm from anything that appeared in the last two hundred years. Computer experiments have been carried out with simulated noisy data, from polynomials of degree 1, 2, 3, and 4, using four well known model selection criteria – AIC, AICc, GIC, and BIC – as well as the proposed method, PTS1, for polynomial degree selection and predictions. The PTS1 is the only one to select the correct degree of a polynomial every time; its performance was always the best. BIC and GIC performances are very similar, and their performances are better than AICc, which performed better than AIC.

The prediction results obtained from the proposed PTS1 are significantly better than those from the existing methods. For both the normalised RMS prediction errors and the normalised variations of RMS prediction errors from the proposed method, their relationships with the values of σ

1) appear to be linear for each of linear, quadratic, cubic, and quartic polynomials,
2) appear to be identical for each of linear, quadratic, cubic, and quartic polynomials, and
3) are described by a single line with a slope of 1.

Therefore, the normalised root-mean square (RMS) prediction errors and the normalised variation of the RMS prediction errors scale linearly with the standard deviation of the noise. As an internal check of the proposed method, two hidden parameters, specifically $< \mu >$ and its RMS, $\sigma_{\mu(q)}$, have been estimated from the data. For each value of the degree of a polynomial and value of the standard deviation of noise, the agreements between estimated values and theoretical values of $< \mu >$ as well as between estimated values and theoretical values of $\sigma_{\mu(q)}$ are remarkable.

ACKNOWLEDGEMENT: The author acknowledges Dr C Liu for formatting the manuscript.

References

[1] A.M. Legendre, "Nouvelles méthodes pour la détermination des orbites des comètes“, Firmin Didot, Paris, 1805. "Sur la Méthode des moindres quarrés" appears as an appendix.

[2] C.F. Gauss, "*Theoria motus corporum coelestium*", Perthes, Hamburg, Germany, 1809. Translation reprinted as "*Theory of motions of heavenly bodies moving about the sun in conic sections*", Dover, New York, 1963.

[3] J D Gergonne, (November 1974) [1815]. "The application of the method of least squares to the interpolation of sequences". Historia Mathematica (Translated by Ralph St. John and S. M. Stigler from the 1815 French ed.), vol. 1,

no. 4, pp. 439–447, 1815 (November 1974), DOI:10.1016/0315-0860(74)90034-2.

[4] S M Stigler, Gergonne's 1815 paper on the design and analysis of polynomial regression experiments". Historia Mathematica, vol. 1, no. 4, pp. 431–439, 1974, DOI:10.1016/0315-0860(74)90033-0.

[5] G U Yule, "On the Theory of Correlation", Journal of the Royal Statistical Society, vol.60, no. 4, pp. 812-854, 1897. DOI: 10.2307/2979746. JSTOR 2979746.

[6] K Pearson, G U Yule, N Blanchard, and A Lee, "The Law of Ancestral heredity", Biometrika, vol. 2, no. 2, pp. 211-236, 1903. DOI:10.1093/biomet/2.2.211. JSTOR 2331683.

[7] R A Fisher, "The goodness of fit of regression formulae, and the distribution of regression coefficients", Journal of the Royal Statistical Society, vol. 85, no. 4, pp. 597-612, 1922. DOI:10.2307/2341124. JSTOR 2341124.

[8] Vaughan Pratt, "Direct least-squares fitting of algebraic surfaces", SIGGRAPH Computer Graphics, vol. 21, no. 4, pp. 145–152, 1987.

[9] A T Kalai, A R Klivans, Y Mansour, and R A Servedio, "Agnostically learning halfspaces", SIAM Journal on Computing, vol. 37, no. 6, pp. 1777–1805, 2008.

[10] Ian B MacNeill, "Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times", The Annals of Statistics, vol. 6, no. 2, pp. 422–433, 1978.

[11] D Kane, S Karmalkar, and E Price, "Robust polynomial regression up to the information theoretic limit", Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science, pp. 391-402, DOI 10.1109/FOCS.2017.43, 2017.

[12] H Linhart and W Zucchini, "Model Selection", New York: Wiley, 1986.

[13] K P Burnham and D R Anderson, "Model Selection and Multi-model Inference", New York: Springer-Verlag, 2002.

[14] H Akaike, "A new look at statistical model identification", IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, DOI:10.1109/TAC.1974.1100705, 1974.

[15] H Akaike, "On the likelihood of a time series model", The Statistician, vol. 27, no. 3, pp. 217-235, 1978.

[16] K P Burnham and D R Anderson, "Multi-model Inference: understanding AIC and BIC in model selection", Sociological Methods & Research, vol. 33, no. 2, pp. 261-304, DOI: 10.1177/0049124104268644, 2004.

[17] H Akaike, "Prediction and entropy", in A C Atkinson and S Fienberg (eds.), A celebration of statistics, Springer, pp. 1-24, 1985.

[18] P S Bandyopadhyay and M R Forster (eds.), "Philosophy of Statistics", North-Holland Publishing, 2011.

[19] A Boisbunon, et al., "Akaike's Information Criterion, C_p and estimators of loss for elliptically symmetric distributions", International Statistical Review, vol. 82, no. 3, pp. 422-439, DOI: doi:10.1111/insr.12052, 2014.

[20] J E Cavanaugh, "Unifying the derivations of the Akaike and corrected Akaike information criteria", Statistics & Probability Letters, vol. 31, no. 2, pp. 201-208, DOI: 10.1016/s0167-7152(96)00128-9, 1997.

[21] G Claeskens and N L Hjort, "Model selection and model averaging", Cambridge University Press, 2008.

[22] J de Leeuw, "Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle", in S Kotz and N L Johnson (eds.), "Breakthroughs in Statistics I", Springer, pp. 599-609, 1992.

[23] N Sugiura, "Further analysis of the data by Akaike's information criterion and the finite corrections", Communications in Statistics – Theory and Methods, vol. 7, pp. 13-26, DOI: 10.1080/03610927808827599, 1978.

[24] C M Hurvich and C L Tsai, "Regression and time series model selection in small samples", Biometrika, vol. 76, no. 2, pp. 297-307, DOI: 10.1093/biomet/76.2.297,1989.

[25] S Konishi and S Kitagwa, Information Criteria and Statistical Modeling, Springer, 2008.

[26] R J Bhansali and D Y Downham, "Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion", Biometrika, vol. 64, no. 3, pp. 547-551, DOI: 10.1093/biomet/64.3.547, 1977.

[27] G Schwarz, "Estimating the dimension of a model", The Annals of Statistics, vol. 6, no. 2, pp. 461-464, 1978.

[28] R I Kashyap, "Optimal choice of AR and MA parts in autoregressive moving average models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 4, no. 2, pp. 99-104, 1982.

[29] J Rissanen, "Modeling by shortest data description", Automatica, vol. 14, no. 5, pp. 465-471, 1978.

[30] J Rissanen, "Estimation of structure by minimum description length", Circuits, Systems and Signal Processing, vol. 1, no. 4, pp. 395-406, 1982.

[31] Y Yang, "Can the strengths of AIC and BIC be shared?", Biometrika, vol. 92, no. 4, pp: 937-950, DOI: 10.1093/biomet/92.4.937, 2005.

[32] S I Vrieze, "Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)", Psychological Methods, vol. 17, no. 2, pp. 228-243, DOI: 10.1037/a0027127, 2012.

[33] P Stoica and Y Selen, "Model-Order Selection", IEEE Signal Processing Magazine, vol. 21, no. 4, pp. 36-47, DOI: 10.1109/MSP.2004.1311138, 2004.

[34] A K Nandi, "Data modelling with polynomial representations and autoregressive representations, and their connections", IEEE Access, Paper ID: Access-2020-26720, DOI: 10.1109/ACCESS.2020.3000860, accepted on 04 June 2020.

[35] C.F. Gauss, "Theoria combinationis observationum erroribus minimis obnoxiae", Henricus Dieterich, Gottingen, Germany, 1823.

Asoke K. Nandi (F'11) received the degree of Ph.D. in Physics from the University of Cambridge (Trinity College), Cambridge (UK). He held academic positions in several universities, including Oxford (UK), Imperial College London (UK), Strathclyde (UK), and Liverpool (UK) as well as the Finland Distinguished Professorship in Jyvaskyla (Finland). In 2013, he moved to Brunel University London (UK), to become the Chair and Head of Electronic and Computer Engineering. Professor Nandi is a Distinguished Visiting Professor at Xi'an Jiaotong University (China) and an Adjunct Professor at University of Calgary (Canada).

In 1983 Professor Nandi co-discovered the three fundamental particles known as W^+, W^- and Z^0 (with the UA1 team at CERN), providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 awarded the prize to his two team leaders for their decisive contributions. His current research interests lie in signal processing and machine learning, with applications to communications, image segmentations, biomedical data, etc. He has made many fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big and Heterogeneous Data".

He has authored over 600 technical publications, including 240 journal papers as well as five books, entitled Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines (Wiley, 2020), Automatic Modulation Classification: Principles, Algorithms and Applications (Wiley, 2015), Integrative Cluster Analysis in Bioinformatics (Wiley, 2015), Blind Estimation Using Higher-Order Statistics (Springer, 1999), and Automatic Modulation Recognition of Communications Signals (Springer, 1996). The H-index of his publications is 75 (Google Scholar) and his ERDOS number is 2.

Professor Nandi is a Fellow of the Royal Academy of Engineering (UK) as well as a Fellow of seven other institutions, including the IEEE and the IET. Among the many awards he received are the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers (UK) in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (UK) in 1998. Professor Nandi is an IEEE EMBS Distinguished Lecturer (2018-19).

Figure Captions

Figure 1 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for a linear polynomial.

The upper half of Figure 2 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), while the lower half of Figure 2 depicts the variation of RMS prediction errors versus the standard deviations of noise for a linear polynomial.

Figure 3 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for a quadratic polynomial.

The upper half of Figure 4 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), while the lower half of Figure 4 depicts the variation of RMS prediction errors versus the standard deviations of noise for a quadratic polynomial.

Figure 5 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations a cubic polynomial.

The upper half of Figure 6 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), while the lower half of Figure 6 depicts the variation of RMS prediction errors versus the standard deviations of noise for a cubic polynomial.

Figure 7 shows the accuracy (%) of degree selection using AIC (green lower triangles), AICc (black upper triangles), GIC (magenta + signs), BIC (blue circles), and PTS1 (red stars) versus noise standard deviations for a quartic polynomial.

The upper half of Figure 8 displays the RMS prediction errors from GIC (magenta + signs), BIC (blue circles), and the proposed method (red stars), while the lower half of Figure 8 depicts the variation of RMS prediction errors versus the standard deviations of noise for a quartic polynomial.

The upper half of Figure 9 displays the normalised RMS prediction errors from the proposed method for linear (red stars), quadratic (red open circles), cubic (red upper triangles), and quartic (red plus signs) polynomials. The lower half of Figure 9 depicts the normalised variations of RMS prediction errors versus the standard deviations of noise. The two blue straight lines have the slope of 1 and the intercept of 0.

All the eight subplots in Figure 10 correspond to $\sigma = 5$. The first column displays histograms of 1,000 values of $< \mu >$ for degree = 1 in the top subplot, for degree = 2 in the next subplot, for degree = 3 in the next subplot, and for degree = 4 in the bottom subplot. The second column displays histogram of 1,000 values of $\sigma_{\mu(q)}$ for degree = 1 in the top subplot, for degree = 2 in the next subplot, for degree = 3 in the next subplot, and for degree = 4 in the bottom subplot.