Performance comparison of model selection criteria by generated experimental data

Radoslav Mavrevski[1,*], *Peter* Milanov[2,3], *Metodi* Traykov[1], *Nevena* Pencheva[4]

[1]Department of Electrical Engineering, Electronics and Automatics, Faculty of Engineering, University Center for Advanced Bioinformatics Research, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., 2700 Blagoevgrad, Bulgaria

[2]Department of Informatics, Faculty of Mathematics and Natural Sciences, University Center for Advanced Bioinformatics Research, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., 2700 Blagoevgrad, Bulgaria

[3]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., Block 8, 1113 Sofia, Bulgaria

[4]Department of Health Cares, Faculty of Public Health, Health Cares and Sports, University Center for Functional Research in Sports and Kinesitherapy, South-West University "Neofit Rilski", 66 Ivan Mihaylov Str., 2700 Blagoevgrad, Bulgaria

Abstract. In Bioinformatics and other areas the model selection is a process of choosing a model from set of candidate models of different classes which will provide the best balance between goodness of fitting of the data and complexity of the model. There are many criteria for evaluation of mathematical models for data fitting. The main objectives of this study are: (1) to fitting artificial experimental data with different models with increasing complexity; (2) to test whether two known criteria as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) can correctly identify the model, used to generate the artificial data and (3) to assess and compare empirically the performance of AIC and BIC.

1. Introduction

Model selection is a process of choosing a model from set of candidate models from different classes, which will provide the best balance between goodness of fitting of the data and complexity of the model [1, 2, 3]. There are different criteria for evaluation of competitive mathematical models for data fitting (approximation). Information criteria provide an attractive base for model selection [1, 3], [4-11]. However, little is understood about their relative performance in model selection.

This research has several specific objectives:

(1) to generate artificial experimental data by known test models;

(2) to fitting data with various models with increasing complexity;

(3) to verify if the class model used to generate the data could be correctly identified through the two commonly used criteria Akaike's information criterion (AIC) and Bayesian information criterion (BIC) and to assess and compare their performance.

2. Materials and methods

2.1 The generate of experimental data

We use the GraphPad Prism software for the artificial experimental data generating and for curve fitting. GraphPad Prism combines nonlinear regression, basic biostatistics, and scientific graphing. (http://www.graphpad.com/scientific-software/prism).
To generate artificial experimental data we use class model - third order polynomial. The individual member of this class is:

$$y = 44 + 99x - 59x^2 + 8x^3 + \varepsilon \qquad (1)$$

$$\varepsilon \sim Normal,\ SD = 5,$$

where ε is random error with Gaussian distribution and standard deviation (SD).

The graph of the third order polynomial:

$a + bx - cx^2 + dx^3$, *a=44, b=99, c=-59, d=8* is shown in Figure 1.

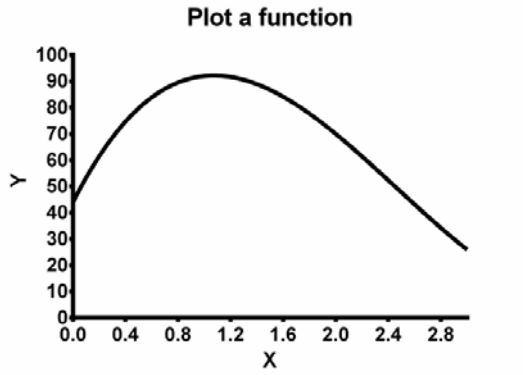

Fig. 1. The individual member of third order polynomial.

* Corresponding author: radoslav_sm@abv.bg

For our computational experiments were generated samples with different sizes - small sample (15 points), middle sample (31 points) and large sample (101 points), following the classification in [12].

2.2 Fitting experimental data

In this research we use different class models (polynomials from first to sixth order) for fitting the artificial experimental data. To find the individual "optimal" models $P^*(M_j)$ in the classes M_j, j = 1, …, 6, we use least squares fitting in GraphPad Prism 6.0. Least squares fitting criterion is defined as follows:

$$F(a) = \sum_{i=1}^{n} (y_i - f(x_i, a_1, \dots, a_s))^2 \qquad (2)$$

The problem is to find $a^* = (a_1^*, \dots, a_s^*)$, such that minimizes $F(a)$.

2.3 Criteria for selection of the optimal model from different class models

2.3.1 Akaike's information criterion

One of the most commonly used criterion for model selection is AIC. The idea of AIC is to select the model that minimizes the negative likelihood penalizing by the number of parameters:

$$AIC = \begin{cases} nln\left(\frac{RSS}{n}\right) + 2k, & \frac{n}{k} \geq 40 \\ nln\left(\frac{RSS}{n}\right) + 2k + \frac{2k(k+1)}{n-k-1}, & \frac{n}{k} < 40, \end{cases} \qquad (3)$$

where n is the number of data points; k is the number of the fitting parameters by the regression plus one (since regression is "an estimating" of the sum-of-squares as well as the values of the parameters); *RSS*, or residual sum of squares, is the sum of the squares of the vertical deviations from each data point to the graph of a curve of the "optimal" fitted model.

2.3.2 Bayesian information criterion

The other most commonly used criterion BIC has the highest posterior probability. AIC and BIC criteria differ only in that the coefficient multiplies the number of parameters. In other words, the criteria differ by how strongly they penalise large models:

$$BIC = nln\left(\frac{RSS}{n}\right) + kln(n), \qquad (4)$$

with the same meaning of *RSS*, n and k, above.

In this situation, the model that minimizes BIC has the highest posterior probability. BIC penalizes the models more from AIC for increasing number of parameters. AIC does not depend directly on the sample size. In general, models chosen by BIC will be more parsimonious than those chosen by AIC.

2.3.3 Program for calculating the AIC and BIC criteria

For calculation of the criteria values of AIC and BIC according to formulas (3) and (4), we use a program "Comparing Models" developed by us in our previous research (see Figure 2), [8].

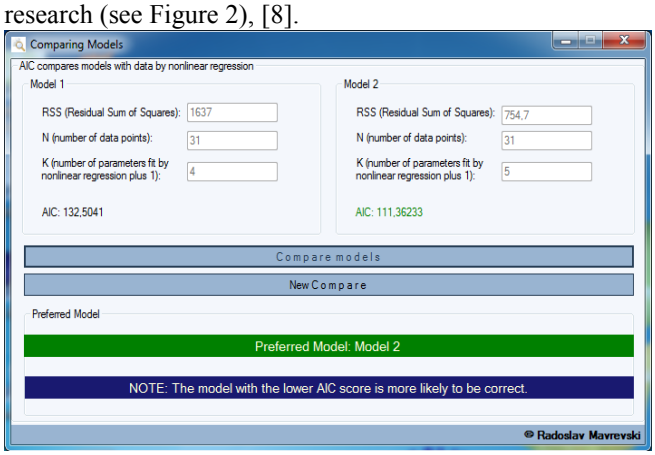

Fig. 2. Example for calculation of the AIC: dialogue box of the program "Comparing Models" for calculating AIC.

3. Results

3.1 Case 1: generate 15 points

For our first experiment we generate 15 points (small sample) in the interval from 0 to 3 with step 0.2. In this case only AIC criterion correctly identified the third order polynomial (true class model) as the optimal model, and BIC criterion chooses fifth order polynomial as the optimal model (false class model) (see Table 1).

Table 1. Result with small sample (15 points).

Polynomial class model	Number of data points	Number of parameters	AIC value	BIC value
First order	15	2	91.99	91.93
Second order		3	67.76	66.59
Third order		4	**60.00**	56.87
Fourth order		5	60.88	54.63
Fifth order		6	64.70	**53.65**
Sixth order		7	74.66	56.32

Simulated data and curves of the fitting models are shown in Figure 3.

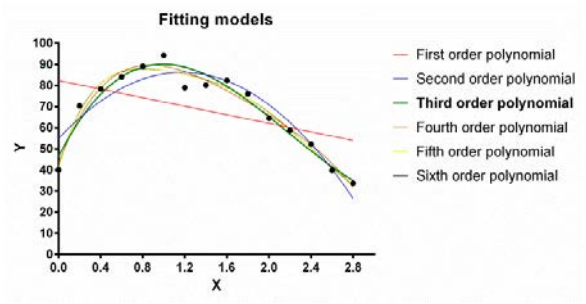

Fig. 3. Simulated data (15 points) and curves of the fitting models (six fitted polynomial curves of increasing order, from 1 (straight line) to 6).

3.2 Case 2: generate 31 points

For the second experiment we use 31 points (middle sample), that are generated in interval from 0 to 3 with step 0.2. Here, both AIC and BIC criteria correctly identified the third order polynomial (true class model) as the optimal model (see Table 2).

Table 2. Result with middle sample (31 points).

Polynomial class model	Number of data points	Number of parameters	AIC value	BIC value
First order	**31**	2	181.15	184.57
Second order		3	132.50	136.70
Third order		4	**111.36**	**116.13**
Fourth order		5	114.46	119.56
Fifth order		6	116.35	121.52
Sixth order		7	119.74	124.67

Simulated data and curves of the fitting models are shown in Figure 4.

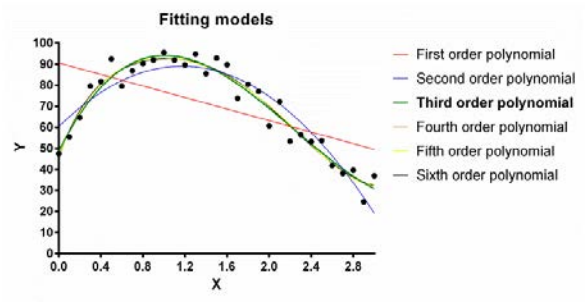

Fig. 4. Simulated data (31 points) and curves of the fitting models (six fitted polynomial curves of increasing order, from 1 (straight line) to 6).

3.3 Case 3: generate 101 points

In the last case we use 101 points (large sample) generated in interval from 0 to 3 with step 0.2. The obtained results showed that in this case only BIC criterion correctly identified the third order polynomial (true class model) as the optimal model, while AIC criterion chooses fifth order polynomial as the optimal model (false class model) (see Table 3).

Table 3. Result with large sample (101 points).

Polynomial class model	Number of data points	Number of parameters	AIC value	BIC value
First order	**101**	2	569.67	577.26
Second order		3	348.03	358.07
Third order		4	303.72	**316.16**
Fourth order		5	305.04	319.84
Fifth order		6	**300.47**	317.57
Sixth order		7	302.01	321.95

Simulated data and curves of the fitting models are shown in Figure 5.

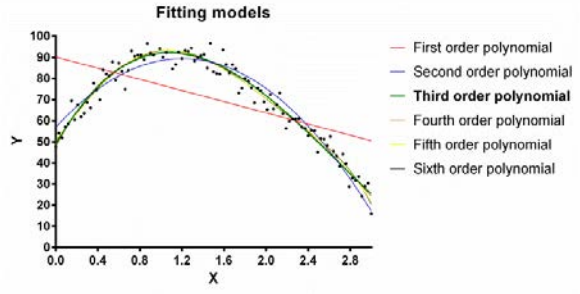

Fig. 5. Simulated data (101 points) and curves of the fitting models (six fitted polynomials curves of increasing order, from 1 (straight line) to 6).

In Figure 6 we show comparison of effectiveness of AIC and BIC in the selection of the optimal model (in all three cases) from the set of 6 class polynomials that was used for fitting the data.

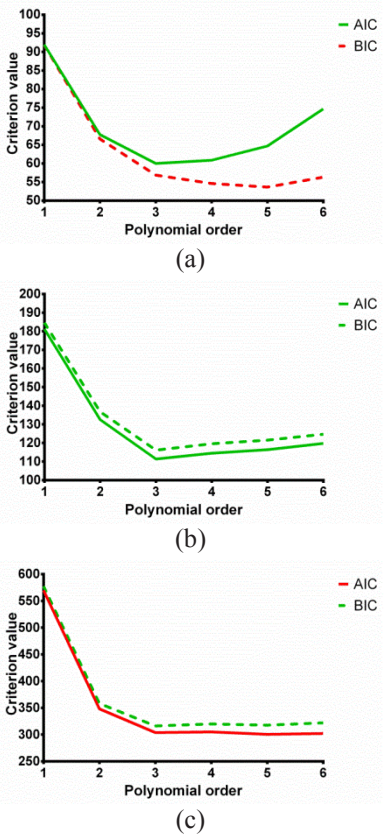

Fig. 6. Comparison of the effectiveness of AIC and BIC: (a) 15 points generated, (b) 31 points generated, (c) 101 points generated.

Figure 6(a) shows that AIC chooses the true class model, and Figure 6(c) shows that BIC chooses the true class model. In Figure 6(b) we can see that both, AIC and BIC, choose the true class model.

4. Discussion

The obtained results from the computational experiments suggested that AIC performs relatively well for small samples but is inconsistent and does not improve performance for large samples. The BIC criterion appears to perform relatively poorly for small samples but is consistent and improves its performance with increasing the sample size. This is consistent with previous studies [4, 13], which demonstrated that BIC is consistent (that is, it tends to choose the true model with a probability equal to 1) in large samples. In our experiments BIC also outperforms AIC when there is a large sample (101 data points) in identification the true class model. As a whole, the current results suggest that generally AIC should be preferred in smaller samples whilst BIC should be preferred in larger samples.

Acknowledgements

This work is partially supported by the project of the Bulgarian National Science Fund, entitled: Bioinformatics research: protein folding, docking and prediction of biological activity, NSF I02 /16 12.12.14.

References

1. H. Acquah, J. Dev. Agric. Econ. **2**, 1-6 (2010)
2. SJ. Ahn, JIPS **4**, 153-158 (2008)
3. H. Akaike, IEEE Trans. Autom. Control **19**, 716–772 (1974)
4. P. Bickel and P. Zhang, J. Am. Stat. Assoc. **87**, 90–97 (1992)
5. P. Burnham and D. Anderson, *Model Selection and Multimodel Inference 2 ed.* (Springer-Verlag, New York, 2002)
6. ID. Coope, J. Optim. Theory. Appl. **76**, 381-388 (1993).
7. B. Joseph and L. Nicole, J. Am. Statist. Assoc. **99**, 279–290 (2004)
8. R. Mavrevski, C. R. Acad. Bulg. Sci. **67**, 1345-1354 (2014).
9. T. Dzimbova, F. Sapundzhi, N. Pencheva, P. Milanov, Int. J. Bioautomation **17**, 5-16 (2013)
10. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Bulg. Chem. Commun. **2,** 613-618 (2015)
11. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Der Pharma Chemica **8**, 118-124 (2016)
12. A. Ghasemi and S. Zahediasl, Int J Endocrinol Metab. **10**, 486-489 (2012)
13. E. Ward, Ecol. Model. **2**, 1-10 (2008).