Assessment of different model selection criteria by generated experimental data

RADOSLAV MAVREVSKI*
Department of Electrical Engineering, Electronics and Automatics
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA
*radoslav_sm@abv.bg https://ais.swu.bg/profile/mavrevski

PETER MILANOV
Department of Informatics, Faculty of Mathematics and Natural Sciences, 3Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
South-West University "Neofit Rilski", Bulgarian Academy of Sciences
66 Ivan Mihaylov Str., 2700 Blagoevgrad, Acad. Georgi Bonchev Str., Block 8, 1113 Sofia
BULGARIA

METODI TRAYKOV
Department of Electrical Engineering, Electronics and Automatics
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA

NEVENA PENCHEVA
Department of Health Cares, Faculty of Public Health
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA

Abstract: - Model selection is a process of choosing a model from a set of candidate models which will provide the best balance between goodness of fit of the data and complexity of the model. Different criteria for evaluation of competitive mathematical models for data fitting have become available. The main objectives of this study are: (1) to generate artificial experimental data by known models; (2) to fit data with various models with increasing complexity; (3) to verify if the 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 empirically their performance. The artificial experimental data generating and the curve fitting is performed through using the GraphPad Prism software.

Key-Words: - model selection criteria, fitting, experimental data

1 Introduction

In the past it was proposed different models, but little work was done on c omparing of the models. Comparing least square errors between measured and modelled data, indicates the quality of fit of each model. However, the least squares statistic does not take into account the tradeoff between model complexity and estimation errors. Models of increased complexity can better adapt to fit to data. However, additional parameters may fit to measurement noise and not describe any important processes. Using solely the model that gives the lowest mean square error will often just lead to the largest model being selected as "optimal". The "optimal" model should balance simplicity and quality of fit.

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 Problem Formulation

There are two main cases of models, which concern experimental data analysis: distribution fitting and curve fitting. Their approaches and purposes lead to two different meanings of the term model [3]. The distribution fitting follows the behavior of a single variable and involves modeling of its probability distribution. This model is in fact a normalized probability density function. The appropriate plot for the experimental data is a histogram. The objective of the analysis of data like these, is not to predict how the value of one variable is related with a value of any other variable, but rather to describe the full frequency distribution of observed variable as a sample of data.

The second case modeling, titled curve fitting is applied when we analyses the behavior of one variable which depend on one or more independent variables and the individual model could be interpreted as a fitting function of the experimental data. Thus, we obtain a curve to a set of points and the appropriate plot for the data is an X-Y scatterplot.

Mathematical models are commonly used in biological sciences. We usually try to determine whether the experimental data are consistent with a particular theoretical relationship and find the model in the class model M describing this relationship. Suppose the experimental data consist of a set of n values of some measured variable y: $y_1, \ldots, y_n$, corresponding to n associated values of some independent variable x: $x_1, \ldots, x_n$. So we obtain data points (x_i, y_i), for $i = 1, \ldots, n$, forming a set of points of the experimental data $A = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. The set of points A can be modeled by a mathematical function $y = f(x;\, p_1, \ldots, p_m)$, with a variable x and parameters $p_1, \ldots, p_m$, using fitting approach [2, 6].

The best fit of the model to the experimental data is quantitatively defined as the minimization of some well described criteria with respect to the parameters of the models. The selection of the "optimal" model describing the experimental data is a topical scientific problem beginning with careful analysis of the experimental data available and going through the following steps: (a) creating a scatter plot of the data and checking whether there is any trend in these data and whether there are among obvious wrong data; if there are obvious erroneous data, then we ignore them and do a gain the measurements or observations; (b) when there is an obvious trend in the set of data, we try to find the class model, let us say polynomials of degree m: $P_m(x, p_s) = p_0 + p_1 * x + p_2 * x^2 + \cdots + p_m * x^m$, for $s = 0, \ldots, m$, polynomial with one variable x and $m+1$ parameters $p_0, \ldots, p_m$, expressing this trend of interrelationships of the studied factors; while observing the cloud with the experimental data, we turn heuristically to classes of models $M_1, \ldots, M_\kappa$ (quadratic, cubic, exponential, hyperbolic and etc.), which describe the set of experimental data in optimal way concerning to minimize the approximation error; (c) after the selection of classes of models, we use different fitting methods for finding the best models in the given classes; finding the individual model in the given classes is usually made by various fitting methods, such as the most widely used method of least squares fitting or other such as robust fitting, minimax fitting etc. [2, 6]; and (d) the selected candidate "optimal" models in each class mentioned above, are compared with some commonly used criteria for the evaluation of these models such as AIC [1, 3, 7, 8, 13] and BIC [1, 3, 7, 8, 13] . The values of these criteria are calculated for each class model and an "optimal" model is selected for each of the fitting methods with respect to the criteria.

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.

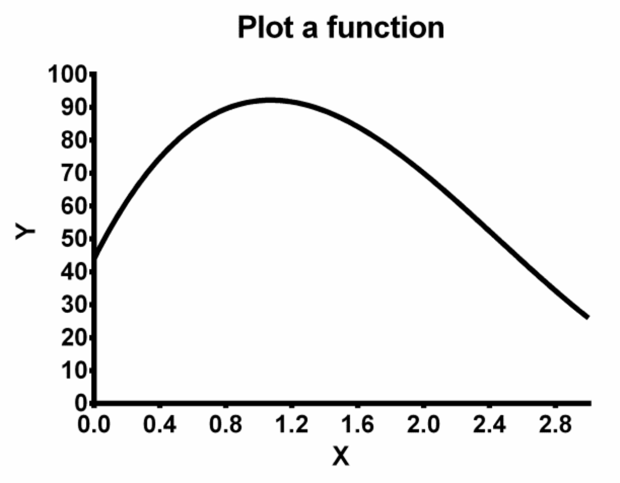


Fig. 1. The individual member of third order polynomial.

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, $j = 1, \ldots, 6$, we use least squares fitting in GraphPad Prism 6.0. Least squares fitting criterion is defined as follows:

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

The problem is to find $a^* = (a_1^*, \ldots, a_s^*)$, 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:

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

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:

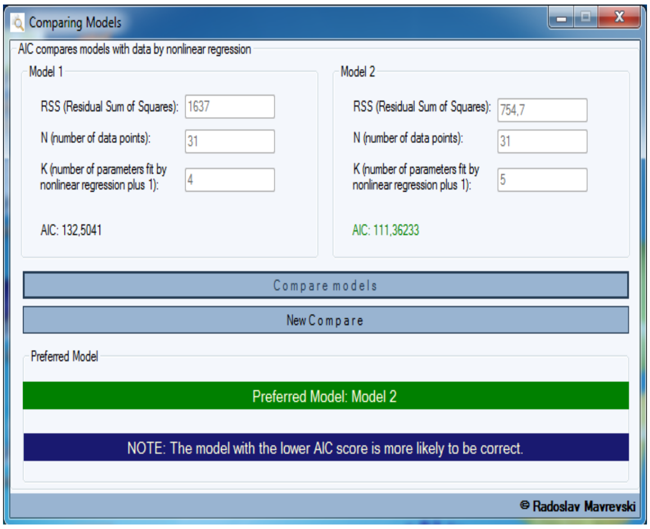
$$BIC = n\ln\left(\frac{RSS}{n}\right) + k\ln(n), \qquad (4)$$

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

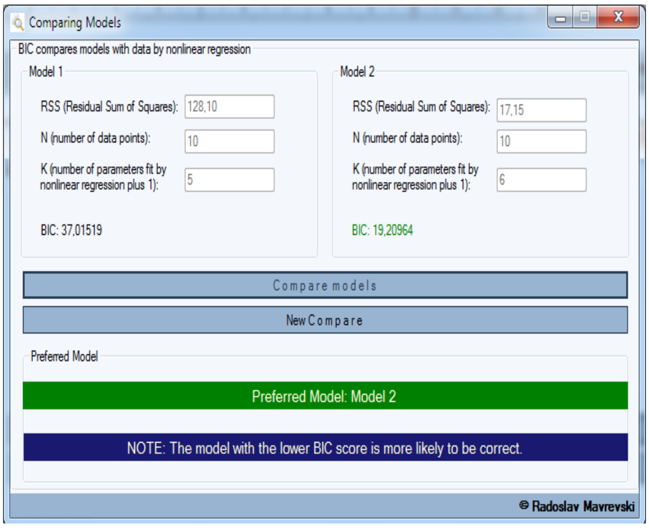
In this situation, the model that minimizes BIC has the highest posterior probability. BIC penalizes the models more from AIC for increasing number of parameters. AIC does not depend directly on the sample size. In general, models chosen by BIC will be more parsimonious than those chosen by AIC.

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].



(a)



(b)

Fig. 2. Example for calculation of the AIC: (a) dialogue box of the program "Comparing Models" for calculating AIC; (b) dialogue box of the program "Comparing Models" for calculating BIC.

3 Problem Solution

3.1 Case 1: generate 15 points

For our first experiment we generate 15 po ints (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.

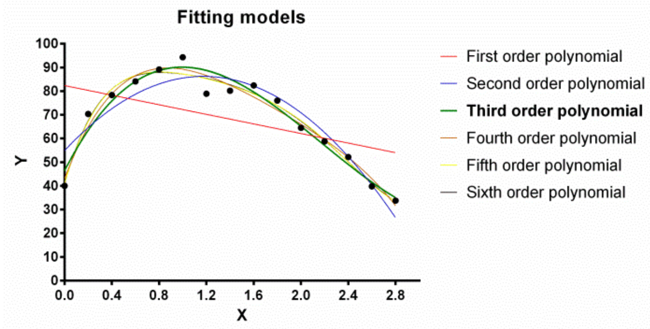


Fig. 3. Simulated data (15 points) and curves of the fitting models (six fitted polynomial curves of increasing order, from 1 (straight line) to 6).

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.

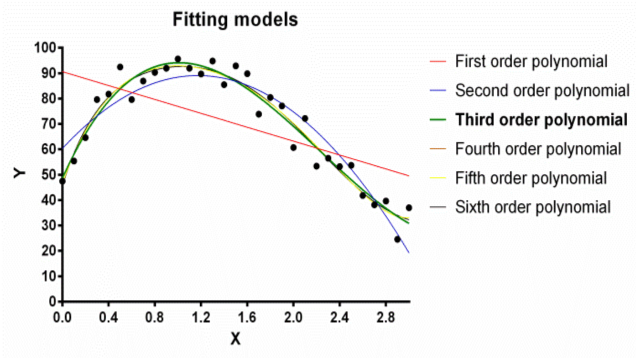


Fig. 4. Simulated data (31 points) and curves of the fitting models (six fitted polynomial curves of increasing order, from 1 (straight line) to 6).

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.

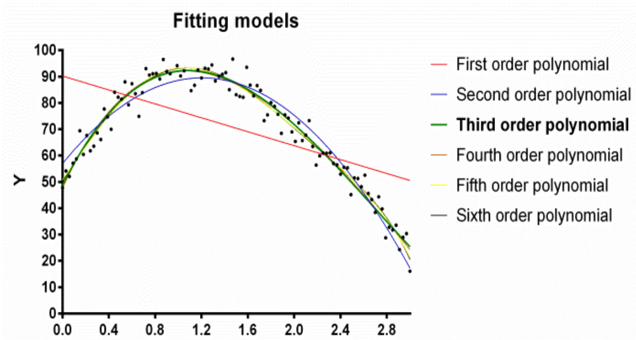
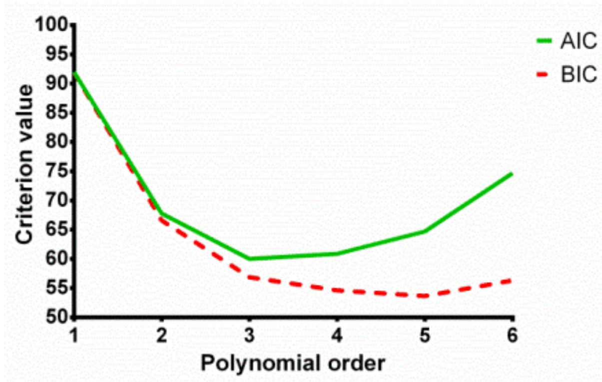
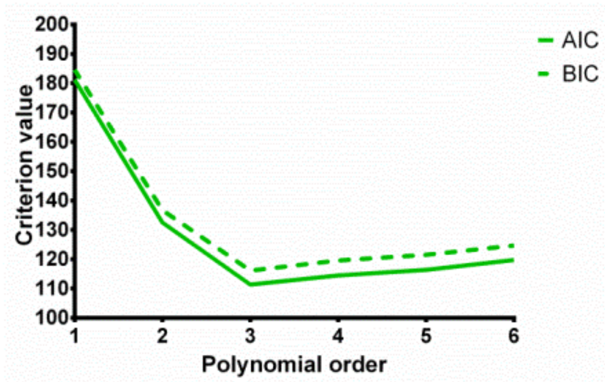


Fig. 5. Simulated data (101 points) and curves of the fitting models (six fitted polynomials curves of increasing order, from 1 (straight line) to 6).

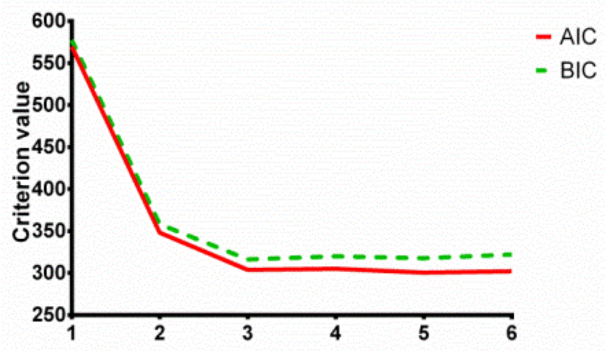
In Figure 6 we show comparison of effectiveness of AIC and BIC in the selection of the optimal model (in all three cases) from the set of 6 class polynomials that was used for fitting the data.



(a)



(b)



(c)

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.

3.4 Coefficient of determination

The coefficient of determination R^2 also can be used to compare regression models. A model with a larger R^2 value means that the independent variables explain a larger percentage of the variation in the independent variable. However, this may conflict with parsimony. Assessment of models in all cases with R^2 are shown in Table 4.

Table 4. Result for R^2 with different sample size.

Polynomial class model	**Number of parameters**	**R^2 value**		
		15 data points	**31 data points**	**101 data points**
First order	2	0.22	0.35	0.34
Second order	3	0.88	0.88	0.93
Third order	4	0.95	0.94	0.95
Fourth order	5	0.96	0.94	0.95
Fifth order	6	**0.97**	**0.95**	**0.96**
Sixth order	7	**0.97**	**0.95**	**0.96**

How to be seen from Table 4 R^2 not a good criterion. Always increase with model size "optimal" is to take the biggest model and never choose the true class model.

3.4 Hausdorff distance used to compare regression models

Hausdorff distance (HD) is a mathematical construct to measure the proximity of two sets of points that are subsets of a metric space and is widely and successfully used in different areas. For example, algorithms for efficiently computing the HD are used for the so-called pattern recognition [14, 15, 16]. An extensive family of applications for the HD may also be found in visualisation, namely in the case of surface remeshing [7, 18]. On the other hand, some authors evaluate different Hausdorff-based algorithms for measuring the similarity between images as p lane sets [19, 20]. The HD is also used as a measure of similarity between two trajectories, represented as sets of points [21].

These findings provoked our interest to investigate the possibility for application of the Hausdorff distance as a cr iterion for "optimal" model selection its effectiveness, comparing it with other model selection criteria such as AIC, BIC and R^2.

In this work we applied an approach in which HD can be applied as a criterion of model selection. HD, between two non-empty sets $A = \{a_1,\ldots, a_n\}$ and $B = \{b_1,\ldots, b_m\}$ in a metric space is defined as follows:

$$\mathrm{HD}(A,B) = \max\big(h(A,B), h(B,A)\big) \quad (5)$$

where:

$$h(A,B) = \max_{a\in A} \min_{b\in B} \|a-b\|, \quad (6)$$

and $\|\bullet\|$ is some underlying norm of A and B (e.g., Euclidean norm). The function h(A, B) is called the directed Hausdorff distance from A to B.

We use the HD as a criterion named Hausdorff distance criterion (HDC) to compare the set of points of the experimental data A and the set of points $B^*(M_j)$, for $j = 1, \ldots, k$, received by the individual "optimal" models $P^*(M_j)$, $j = 1, ..., k$. We are interested in finding $B^* \in \{B_1, \ldots, B_k\}$ such that $\mathrm{HDC}(A, B^*) = \min \mathrm{HD}(A, B^*(M_j)), j = 1, \ldots, k$.

In this study was applied the following algorithm for calculating the HD presented in pseudocode:

input: *Set of points A, Set of points $B^*(M_j)$, $j = 1, \ldots, k$, $i = 1, \ldots, 6$.*

1. For any point a in A finding minimum distance to any point b_i in $B^*(M_j)$;

2. Finding the point a, in A with the largest minimum distance to any point in $B^*(M_j)$;

3. For any point b in $B^*(M_j)$ finding minimum distance to any point a_i in A;

4. Finding the point b, in $B^*(M_j)$ with the largest minimum distance to any point in A;

5. Finding the largest distance between directed Hausdorff distance h(A, $B^*(M_j)$) and directed

Hausdorff distance $h(B^*(M_j), A)$ (see formula (2)).
output: *HD between A and $B^*(M_j)$.*

We have also developed a separate module for the Comparing Models program, which calculates the HDC (see Figure 7).

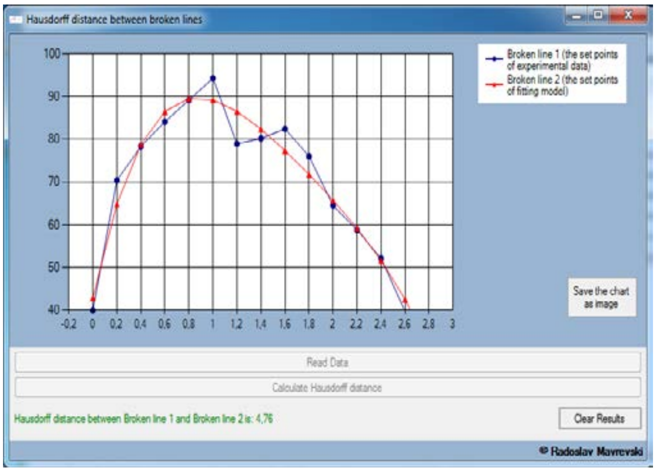
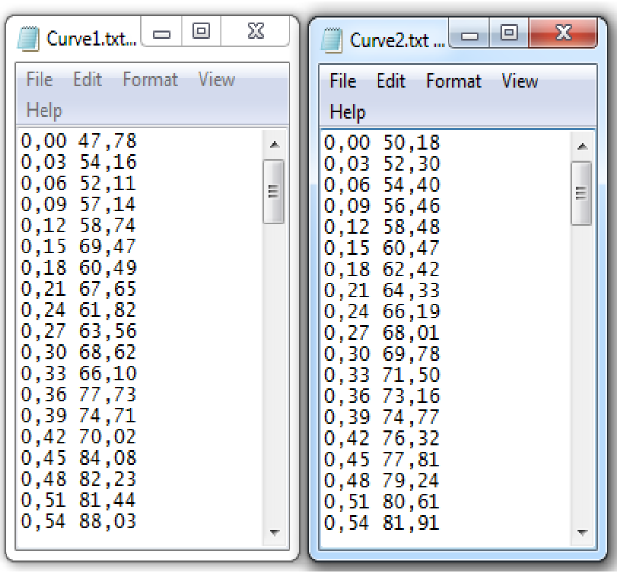


Fig. 7. Example for calculation of the HDC: dialogue box of the program "Comparing Models" for calculating HDC.

As input of the module for calculating the HDC is given the set of points of the experimental data and the set of points, received by the according model (see Figure 8).



(a) (b)

Figure 8. Both input files (101 points) of the module for calculating the HDC: (a) the set of points of the experimental data; (b) the set of points, received by the sixth order polynomial.

This module of the program works only with two text files, containing the two sets of points respectively the A and B and have possibility to display graphically the two sets of points A and B and to constructs the polylines between points, as and ability to export a graph in TIFF format (see Figure 9).

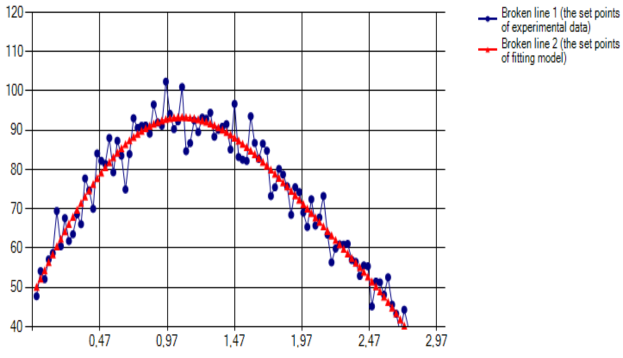


Figure 9. Graph of the sixth order polynomial (101 points), including the value of HD, in TIFF format.

Assessment of models in all cases (15, 31 and 101 points) with HDC are shown in Table 5.

Table 5. Result for HDC with different sample size.

Polynomial class model	Number of parameters	HDC		
		15 data points	31 data points	101 data points
First order	2	20.26	24.98	34.48
Second order	3	8.06	6.56	12.82
Third order	4	6.28	**6.30**	10.04
Fourth order	5	**4.76**	6.38	10.18
Fifth order	6	6.22	8.42	**8.93**
Sixth order	7	6.42	8.20	8.97

Results from Table 5. show that HDC correctly identified the third order polynomial (true class model) as the optimal model in case with middle sample (31 points) and not identified the optimal model, when we have small and large sample.

Usually, in optimization packages, the algorithms for finding the optimal solution are of gradient type and there is not always a guarantee that we have found the global minimum. That is why we solve the fitting problem many times, by starting the gradient method with points which are far or very close to the solution found at the beginning. If the application of this procedure leads us in the solution initially found, then it can be argued with a high probability that the solution found (in respective case of the parameters of the polynomials function) is the global minimum [8].

4 Conclusion

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 w hole, the current results suggest that generally AIC should be preferred in smaller samples whilst BIC should be preferred in larger samples.

In the general case if you want to select the "optimal" model with the smallest mean square error, the AIC and / or BIC criteria are very appropriate. They will choose the optimal model that has the relatively same error for each point in the experimental data point. These criteria report a compromise between the complexity of the model (number of parameters) and the accuracy [1, 3, 7].

The comparison of interpolation utility of regression models with a different number of parameters cannot be done by simple comparison of R^2. At least the adjusted R^2 must be used, but more sophisticated measures like an AIC, BIC are strongly advised.

The HD was successfully applied as a cr iterion for optimal model selection. An efficient algorithm was presented in this study to compute the HD between set of points of the experimental data and set of points of the fitting model. The numerical results obtained in this research are so encouraging that we believe that it may be used in other applications, as well. For checking if the introduced by us criterion for model selection HDC fulfills its intended purpose, we made a comparison of the results obtained by this criterion with the results of the other commonly used criteria for optimality models such as AIC and BIC.

However, there is one problem that may arise when using Hausdorff distance for determining the proximity between two polylines A and B. In this case the two polylines A and B are represented as two sets of points $\{a_1, ..., a_n\}$ and $\{b_1, ..., b_n\}$, respectively and the HD yields a distance which is not indicative for the two polylines in general. One approach to mitigate this problem, proposed in [17] is for the sampling rate of the polylines to be greater, i.e two sets A and B to contain a g reater number of points. In our case, when we incrase sample from 15 t o 31 points HDC correctly identified the third order polynomial (true class model). However, with continued increase in sample to 101 points HDC not identified the optimal model. This is probably due to the fact that it does not include punishment to increase the number of parameters in the model In contrast to AIC and BIC.

A disadvantage of the used polynomial models is that in many cases they are not suitable for the biological interpretation and are worthless outside the range of observed data, i.e. cannot be used to predictions beyond this range.

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, Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship, J. Dev. Agric. Econ., **2**, 2010, 1-6.

[2] SJ. Ahn, Geometric Fitting of Parametric Curves and Surfaces, JIPS **4**, 2008, 153-158.

[3] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, **19**, 1974, 716–772.

[4] P. Bickel and P. Zhang, Variable Selection in Nonparametric Regression with Categorical Covariates, J. Am. Stat. Assoc. **87**, 1992, 90–97.

[5] P. Burnham and D. Anderson, *Model Selection and Multimodel Inference 2 ed.*, Springer-Verlag, New York, 2002,

[6] ID. Coope, Circle fitting by linear and nonlinear least squares, J. Optim. Theory. Appl. **76**, 1993, 381-388.

[7] B. Joseph and L. Nicole, Methods and Criteria for Model Selection, J. Am. Statist. Assoc. **99**, 2004, 279–290.

[8] R. Mavrevski, Selection and comparison of regression models: estimation of torque-angle relationships, C. R. Acad. Bulg. Sci. **67**, 2014, 1345-1354.

[9] T. Dzimbova, F. Sapundzhi, N. Pencheva, P. Milanov, Computer modeling of human delta opioid receptor, Int. J. Bioautomation **17**, 2013, 5-16.

[10] F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Exploring the interactions of enkephalin and dalargin analogues with the mu-opioid receptor, Bulg. Chem. Commun. **2,** 2015**,** 613-618.

[11] F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Comparative evaluation of four scoring functions with three models of delta opioid receptor using molecular docking, Der Pharma Chemica **8**, 2016, 118-124.

[12] A. Ghasemi and S. Zahediasl, Normality Tests for Statistical Analysis: A Guide for Non-Statisticians, Int J Endocrinol Metab. **10**, 2002, 486-489.

[13] E. Ward, Ecol. Model. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools, **2**, 2008, 1-10.

[14] P. Daniel, A. Gregory, J. William, Comparing Images Using the Hausdorff Distance, IEEE Trans. Pattern Anal. Mach. Intell., **15** 1993, 850–863.

[15] C. DU, G. SU, Improved line segment Hausdorff distance for face recognition, J. Optoelectronics Laser, **16**, 2005, 89–93.

[16] S. GAO, M.K. LEUNG, Face recognition using line edge map, IEEE Trans. Pattern Anal. Mach. Intell., **24**, 2002, 764–779.

[17] M. Barton, I. Hanniel, G. Elber, M. Kim, Precise Hausdorff Distance Computation between Polygonal Meshes, Comput. Aided Geo. Des., **27**, 2010, 580–591.

[18] Manocha D., C. Erikson, Gaps: General and automatic polygonal simplification, Proc. of the ACM Symposium on Interactive 3D Graphics, 1998, 79–88.

[19] Shapiro M., M. Blaschko, On Hausdorff Distance Measures, University of Massachusetts Amherst, Computer Science Dept., Technical report, 2004.

[20] Son J., Kim S.H., J.S. Kim, Text image matching without language model using a Hausdorff distance, Inform. Process Manag., 44, 2008, 1189–1200.

[21] C. Tan, Y.J. Zhang, Computing eigenface from edge images for face recognition based on Hausdorff distance, Proc. of the 4th International Conference on Image and Graphics, ICIG 2007, 639–644.