African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 5, Issue 3, 2026
Page No: 345-351
Website: *https://aaasjournals.com/index.php/ajapas/index*
ISI 2025: 1.126 **SJIFactor 2024: 6.752** **معامل التأثير العربي: 1.62**

Semi-Parametric Model Selection Using AIC and BIC under the Curse of Dimensionality: A Monte Carlo Investigation

Nawal Mohamed Othman Mustafa *
Department of Statistics, Faculty of Arts and Sciences, Al-Marj, University of Benghazi, Al-Marj, Libya

اختيار النموذج شبه المعلمية باستخدام معيار معلومات أكايكي (*AIC*) ومعيار معلومات بايز (*BIC*) في ظل مشكلة الأبعاد: دراسة باستخدام طريقة مونت كارلو

نوال محمد عثمان مصطفي *
قسم الإحصاء، كلية الآداب والعلوم المرج، جامعة بنغازي، المرج، ليبيا

**Corresponding author:* *nawal.mustafa@uob.edu.ly*

Received: May 22, 2026	Accepted: August 14, 2026	Published: August 23, 2026

Abstract:
This article investigates the performance of two common information criteria, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) in choosing among three semiparametric regression models, i.e., the single-index model (SIM), the partial linear model (PLM), and the additive model (AM). The paper also includes a Monte Carlo experiment for various sample sizes and dimensions (n = 50, 100, 200) and (p = 10, 30). According to AIC and BIC, the best model for each simulated data set is the model with the minimum value of the information criterion. CMSR considers the model selection-based efficiency of AIC and BIC, while APE measures predictive performance. AIC and BIC are functions of the sample size and the dimensionality but versus the simulation results of the competing semiparametric models, they seem to behave antithetically for certain competing semi parametric models on the page. However the precision of both criteria can be further increased in model selection and prediction by a larger sample size. Results further indicate that for the increasing sample size, the PLM produces the least prediction error when compared with the SIM and the AM

Keywords: Curse of Dimensionality; Semiparametric Regression; Akaike Information Criterion (AIC); Bayesian Information Criterion (BIC); Model Selection; Single-Index Model (SIM); Partial Linear Model (PLM); Additive Model (AM).

الملخص
تبحث الدراسة الحالية في أداء معياري معلومات أكايكي (AIC) والمعلومات البايزي (BIC) في اختيار نماذج الانحدار شبه المعلمية، وذلك من خلال دراسة محاكاة مونت كارلو بأحجام عينات مختلفة ((n=50,100,200)) وأبعاد نموذجية متزايدة ((p=10,30)). وشملت الدراسة ثلاثة نماذج هي: نموذج المؤشر المفرد (SIM)، والنموذج الخطي الجزئي (PLM)، والنموذج الجمعي (AM). وتم تقييم أداء المعيارين باستخدام معدل اختيار النموذج الصحيح (CMSR) ومتوسط خطأ التنبؤ (APE)، مع دراسة تأثير حجم العينة ونوع دالة الانحدار وزيادة عدد المتغيرات في كفاءة AIC وBIC في اختيار النموذج المناسب.

الكلمات المفتاحية: لعنة الأبعاد؛ الانحدار شبه المعلمي؛ معيار معلومات أكايكي (AIC)؛ معيار المعلومات البايزي (BIC)؛ اختيار النموذج؛ نموذج المؤشر الأحادي (SIM)؛ النموذج الخطي الجزئي (PLM)؛ النموذج الجمعي (AM).

Introduction

A variety of statistical models have been established to account for the changeable nature of data, and to more flexibly model linear and nonlinear relationships than traditional linear models can. However, the burgeoning sample size and dimensionality of observed variables in practice have made model selection more difficult for statisticians. This increasing complexity in model building has much to do with the phenomenon referred to as the "curse of dimensionality." In order to overcome this problem, quasi-parametric models have been proposed as flexible methods that can incorporate both advantages of parametric and nonparametric modeling. Thus, examining the issue of model selection under the curse of dimensionality is the focus of this paper via an exhaustive Monte Carlo analysis, with special emphasis on the comparative performance of Akaike's and the Bayesian Information Criteria (AIC and BIC) in the selection of quasi-parametric models. More specifically, the paper investigates the extent to which each criterion correctly identifies the true model for different sample sizes, number of explanatory variables, and model complexity. The general goal is to learn under which conditions AIC and BIC achieve the best model-selection performance to advance recommendations about which information criterion to use to analyze high-dimensional data.

. 1.1The Curse of Dimensionality

The essence of the curse of dimensionality is that, as the number of explanatory variables increases, the complexity of the model increases, and it becomes almost impossible to graphically display or interpret the results. In addition, it greatly increases the amount of data needed to maintain the precision of the estimates at an acceptable level. For this reason, the curse of dimensionality is one of the fundamental difficulties that has motivated the development of methods for reducing dimensionality and the use of quasi-parametric models, such as the single-index model, the partial linear model, and additive models, with the aim of reducing the effective dimensionality of the problem while maintaining an appropriate degree of flexibility in the relationship (Bellman, 1961; Hastie et al., 2009; Härdle et al., 2004; Härdle & Stoker, 1989; Ichimura, 1993).

1.2 Semi-Parametric Regression Models

Semi parametric regression is a statistical method that blends parametric and nonparametric elements to form a single model. One component of the predictor-response relation is supposed to follow a known functional form that depends on a finite-dimensional parameter, and the other component is unspecified and estimated via nonparametric methods. In this sense, semi parametric regression is a trade-off between the parametric regression model and the nonparametric regression model it enables one to investigate more complex and nonlinear relationships between variables(Ruppert et al., 2003).

1.3 Partial Linear Model – PLM

The Partially Linear Model combines a linear regression model with unknown parameters and a non-parametric model with an unknown function. This model enables one to see a linear dependence of the dependent variable on a certain subset of explanatory variables and a nonlinear dependence of the dependent variable on the rest of explanatory variables, without the prior definition of the functional form of these dependencies. Hence, strike a nice intermediate to combine the interpretability of the linear part with the flexibility of the nonparametric part. This is especially important when one considers that for complex relationships, traditional linear models (or fully non- parametric models of the respondent or predictor variables) often turn out to be inadequate to seize the subtle details and complicated properties of the data.

The general form of the (PLM) model is

$$y = XB + f(Z) + \varepsilon \quad (1)$$

y is denoted the change in dependent variable, where X represents the design matrix of independent variables. The function $f(z) \rightarrow$ represents smoothing function with random vector z, and error term is normally distributed with mean zero and constant variance $\varepsilon \sim N(0, \sigma)$

1.4 Semi parametric Single Index Model – SIM

is estimated non-parametrically The intuition behind the model is to compress a group of multidimensional explanatory variables into one linear index and then permit the index-dependent variable relationship to be an unknown function which (Ichimura, 1993; Carroll et al., 1997).

The one-index model in the case of a continuous dependent variable can be expressed by the formula:

$$y_i = m\left(X_i^{\,T} B\right) + \varepsilon_i \quad \text{i=1,2,…,n} \quad (2)$$

y_i Vector of explanatory variables

$X_i = \left(x_{i1}, x_{2i}, \cdots, x_{ip}\right)^T$:- The vector of model parameters—the parametric component of the model.

$B = (B_1 B_2, \cdots, B_P)^T$ **:** The one index that reduces the explanatory variables to a single variable.

$X_i^T B$; An unknown function representing the non-parametric part of the model.

$m(\cdot)$ The random error term; it is often assumed that $E(\varepsilon_i / x_i) = 0$.

Therefore the model is based on the assumption that the influence of the explanatory variables on y is not a direct function of all variables $x_1, \cdots, x_p$ but a function of a single linear combination through which

$$T_i = X_I^T B$$

Then, the relationship between Y_i and T_i is estimated using the unknown function $m(T_i)$.

[1] **1.5 Additive Model**

The main advantage of the additive model is that it can be used to estimate the effects of each explanatory variable on the dependent variable, separately The model is more flexible than the usual linear regression model and this helps alleviate some of the estimation difficulties encountered in multi-dimensional non-parametric regression. Different smoothing procedures (e.g., kernels, splines) may be used to estimate additive functions and smoothing parameters are employed to achieve an optimal balance between data fit and function smoothness (Wood, 2017).

Math formula

$$y_. = \alpha + \sum_{j=1}^{p} g_j(x_{ij}) + \varepsilon_i$$

where each variable is allowed to influence Y through its own specific non-linear function.

1.6 Akaike Information Criteria, AIC

The Akaike information criterion (AIC) is an established model selection tool that takes into account the goodness of fit of the statistical model and complexity of the model. Following Akaike's (1974) proposal, AIC is based on information theory. The AIC seeks to find a model that achieves a good balance between goodness of fit and predictive quality in terms of new (unseen) data. It adjusts for model complexity by adding a function of the number of parameters estimated to the value of the criterion. The AIC can be expressed as:

$$ALS = -2\log L(\hat{\theta}) + 2K$$

K **:** The amount of parameters (predictors) included in the model

$L(\hat{\theta})$: maximum likelihood function

1.7 Bayesian Information Criterion, BIC

the number of parameters and the sample size. The BIC is most often defined as: model. Instead, BIC applies a penalty term that is a function of The BIC is yet another commonly used model selection criteria. It was introduced by Schwarz (1978) and, similarly to AIC, it trades off goodness of fit of the model with complexity of the

$$BLC = -2\log L(\hat{\theta}) + K\log(n)$$

$L(\hat{\theta})$: It represents the maximum likelihood function of the model.

K ; The number of parameters estimated in the model

n : Sample size

The first term measures the misfit of the model, and the second is a penalty for model complexity. Since the penalty term grows with the number of parameters as well as the sample size, BIC tends to prefer more parsimonious models than AIC, especially for larger samples. The model with the lowest BIC is the preferred model among a set of competing models estimated on the same data.

1.8 Model Selection and Prediction Performance Measures

Correct Model Selection Rate (CMSR)

The model selection rate is a metric used in simulation studies to assess how well a model selection criterion can select the correct model from among a group of competing models.

It is calculated

$$CMSR = \frac{\sum (\hat{M}_i = M_0)}{N} \chi 100$$

N **:** Number of Monte Carlo simulation iterations.

M_0 : The true model used to generate the data.

$\hat{M}_i$: The model chosen in iteration i using AIC or BIC.

I($\hat{M}_i = M_0$): The indicator function, which takes the value...

I=1 If the correct model is selected
I=0 The correct model was not selected.

Average Prediction Error (APE)

It is used to assess the accuracy of a given model's prediction, and calculates the average difference of the actual values with respect to the predicted values .If we have n observations then the formula is

$$APE = \frac{1}{n} \sum_{I=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

where:

Y_i **:** The true value of watching i

$\hat{Y}_i$: The value predicted by the model

n : Sample size.

2.The Simulation Study

A simulation study was conducted on the behavior of the AIC and BIC norms by comparing three semi-parametric regression models, the single-index model (SIM), the partial linear model (PLM), and the summation model (AM), under various high dimensional data setting. In order for the model selection performance to be tested when facing dimensionality difficulties, two sizes of explanatory variables p = 10 and 30, and three sizes of samples (n = 50, 100 and 200) were considered The simulations were done in the R language, and the correctness of model selection rate (CMSR) and the average prediction error (APE) were adopted in the evaluation. The Monte Carlo method was used to repeat each experimental condition 500 times, generating independent training and test samples in each run. The three models were fitted and for each the AIC and the BIC were computed. The model with minimum norm value was then chosen and its prediction error was used as a measure of its performance.

Table1: The Examples some Regression Functions used simulation study [2]

Example	$f_i(t)$		
1)	$\sin(x^2) + 0.5x$	Sample size n	**50,100.200**
2)	$\cos(2x) + 0.5x^2$	Explanatory variables	**P=10,30**
		Model selection criterion	**AIC, BIC**
		Model selection evaluation criterion	**CMSR**
		Forecast evaluation standard	**APE**

2.1 Numerical Summary for the simulation study

are given in Tables 2-5. These tables display a complete description of all results for SIM, PLM and AM with respect to AIC and BIC. The two measures for the competing models under consideration are the Correct Model Selection Rate (CMSR) - How often AIC and BIC select the true model; and the Average Prediction Error (APE) - The predictive performances of the competing models. For a variety of sample sizes and numbers of covariates, we present the APE values in Tables report the results for a range of sample sizes and numbers of covariates. Smaller APEs imply better predictive performance and larger CMSR imply the better model selection capability.

Table 1.2. Correct Model Selection Rates (CMSR, %) for PLM, SIM, and AM under AIC and BIC for different sample sizes and numbers of covariates based on 500 Monte Carlo replications using the first regression function

		n	PLM		SIM		AM	
			P=10	P=30	P=10	P=30	P=10	P=30
CMSR	AIC	50	73.6	76.0	68.6	74.0	99.6	98.8
		100	91.0	92.2	88.6	88.6	100.0	100.0
		200	93.6	94.8	96.0	95.0	100.0	100.0
	BIC	50	98.0	98.4	92.8	95.2	52.6	51.4
		100	100.0	100.0	99.0	99.2	78.0	77.8
		200	100.0	100.0	99.8	100.0	98.6	99.0

Table1.3. Average Prediction Error (APE) for PLM, SIM, and AM under AIC and BIC for different sample sizes and numbers of covariates based on 500 Monte Carlo replications using the first regression function**.**

		n	PLM		SIM		AM	
			P=10	P=30	P=10	P=30	P=10	P=30
APE	AIC	50	0.6000960	0.5671561	0.963121	0.9628836	1.7341605	1.7503481
		100	0.3986215	0.3913126	0.637687	0.6454536	1.0440707	1.0276394
		200	0.3646582	0.3591040	0.591870	0.5935462	0.8608828	0.8658813
	BIC	50	0.4435962	0.4473818	0.723056	0.7355223	1.6840472	1.6298697
		100	0.3820081	0.3762523	0.615677	0.6177070	1.1111791	1.0954668
		200	0.3602726	0.3552347	0.587291	0.5880902	0.8674419	0.8696276

Table1.4. Correct Model Selection Rates (CMSR, %) for PLM, SIM, and AM under AIC and BIC for p=10 and p=30 and different sample sizes based on 500 simulation replications using the second regression function.

		n	PLM		SIM		AM	
			P=10	P=30	P=10	P=30	P=10	P=30
CMSR	AIC	50	74.8	78.2	96.8	98.4	100.0	100.0
		100	92.2	91.8	100.0	100.0	100.0	100.0
		200	92.4	94.0	100.0	100.0	100.0	100.0
	BIC	50	99.6	100.0	100.0	100.0	94.8	94.4
		100	100.0	100.0	100.0	100.0	100.0	100.0
		200	100.0	100.0	100.0	100.0	100.0	100.0

Table1.5. Average Prediction Error (APE) for PLM, SIM, and AM under AIC and BIC for different sample sizes and numbers of covariates based on 500 simulation replications using the second regression function.

		n	PLM		SIM		AM	
			P=10	P=30	P=10	P=30	P=10	P=30
APE	AIC	50	3.4949777	3.0906789	10.724254	13.15857	13.189149	11.575616
		100	0.8202515	0.7198498	2.5981248	3.141657	3.5541999	3.6820370
		200	0.5638714	0.5505692	1.0450076	0.969625	2.0183713	2.0585086
	BIC	50	1.4571990	1.5726297	9.5582051	12.41095	12.944578	11.111019
		100	0.7526503	0.6532133	2.5981248	3.141657	3.5541999	3.6820370
		200	0.5424902	0.5383969	1.0450076	0.969625	2.0183713	2.0585086

Analysis of Simulation Results

The content of Tables 1.2 and 1.3 are analyzed to show the effect of two values for p, 10 and 30, on the performance of the three models, PLM, SIM and AM, when using the AIC and BIC criteria, with sample sizes of n = 50, 100 and 200, for the first regression function based on 500 replications of the Monte Carlo.

[3] 1-The results show that CMSR generally improves as the sample size increases, indicating an improvement in the ability of model selection criteria to identify the correct model.

[4] 2- AM was preferred the most when applying using the AIC; the rate of correct selection was above 98% with n = 50 and was 100% with n = 100 and n = 200. The performance of both PLM and SIM increased with the sample size as well.

[5] 3- Regarding BIC, it performed best in choosing the PLM and the SIM models, with the CMSR increasing to 100% for large samples. Conversely, the performance of the BIC for the AM model was poor at n=50, but increased substantially with larger sample size and got close to 99% at n=200.

Table 1.3 shows that the predictive capabilities of the three model under the AIC and the BIC were very different from each other in a clear way. There is a general decreasing trend for the APE values between the sample size n=50 and n=200 due to better predictive performance by the models when a larger sample size is used.

- When AIC b is applied, the PLM model has the smallest APE values at almost all sample sizes, for not only p=10 but also p=30. At p=10, the APE is 0.6001 for n=50 in the PLM model, and it decreases to 0.3986 for n=100 and 0.3647 for n=200. In a similar manner at p=30, the values are 0.5672, 0.3913 and 0.3591 descending. This suggests that a significant enhancement in the predictive performance of PLM can be achieved when the size of samples gets larger.
- For the SIM model, the APE values were also larger than the PLM model values in every case. At p=10, the APE value decreased from 0.9631 at n=50 to 0.5919 at n=200, and at p=30, it decreased from 0.9629 to 0.5935. This suggests that the SIM model is worse than the PLM model in prediction for this two cases.
- In contrast, the AM model produced the largest APE values among the other two models. Its values at n=50 were about 1.7342 and 1.7503 for p=10 and p=30, respectively, and then they declined to 0.8609 and 0.8659 at n=200 Therefore, the AM model was the least efficient considering the predictive accuracy among the family of APE based models.
- For the BIC, the same overall pattern of results was observed: the PLM resulted in the best predictive performance, followed by the SIM, and the AM suffered from worst predictive errors. In particular, for n=200, the PLM obtained APE values of 0.3603 and 0.3552 when p=10 and p=30, respectively; when compared with those of the SIM of 0.5873 and 0.5881, and the AM of 0.8674 and 0.8696.

the sample size n

As the sample size n grows, the values of the APE are getting smaller for all three models, implying that they become more predictive accurate; this improvement with sample size was especially remarkable for the AM model.

the Influence of the Number of Variables p

The impact of varying the number of variables from p=10 to p=30 was modest for the PLM and SIM models, while the AM model still produced the largest APE values in the majority of cases.

Results of the Tables 1.4 and 1.5 PLM, SIM, and AM — AIC and BIC at p=10 and p=30 with the sample sizes of n=50, 100, and 200 for the three models using the 500 runs with the second example regression function.

- The simulation results indicated that the CMSR increased with the sample size and the selection rates converge to 100% at (n=100) and (n=200). BIC was observed to perform better than AIC in terms of choosing the correct model, especially for the small sample size of (n=50) where the superiority of BIC was more significant for PLM. When the sample size grew, the difference between AIC and BIC became smaller with both criteria having very high rates of correct model selection.
- The APE results decreased obviously as the sample size increased, which shows that with the larger samples, better predictive accuracies were obtained. Out of the three models, PLM usually reached the best (smallest) APE, indicating better predictive performance than SIM and AM in the considered simulation scenarios. This superiority was especially pronounced for the smaller sample size of (n=50).
- In general the simulation results suggest that larger sample size leads to better results for model selection as well as for prediction. BIC showed better model-selection properties, also for small samples, while PLM attained the best predictive performance in terms of smallest APE. These results may indicate that model-selection efficiency and predictive accuracy may favor two different components of model assessment: BIC performs well in detecting the true model whereas PLM performs better for prediction under the second regression function.

Conclusion

The conclusion from the study is that the effectiveness of the use of AIC and BIC selection criteria depends on the size of the sample and the number of explanatory variables. The simulations showed that an increase in sample size resulted in a higher CMSR and lower APE. It means that there are improvements in the efficiency of model selection and predictions respectively. In addition, it was found that BIC is more efficient in terms of choosing the correct model, especially with a small sample size. In terms of making predictions, PLM showed the highest accuracy according to the APE criterion, then SIM and then AM. Therefore, the study suggests using BIC when the priority is selecting the correct model and PLM when making predictions, but it is important to mention that the sample size should be increased**.**

Compliance with ethical standards

Disclosure of conflict of interest

The author(s) declare that they have no conflict of interest.

References

1- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

2- Aneiros-Pérez, G., González-Manteiga, W., & Vieu, P. (2004). Estimation and testing in a partial linear regression under long-memory dependence. *Bernoulli, 10*(1), 49–78.

3- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). John Wiley & Sons.

4-.Härdle, W., Liang, H., & Gao, J. (2000). *Partially linear models*. Springer.

5-Härdle, W., & Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association, 84*(408), 986–995.

6-Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall.

7-Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

8-Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics, 58*(1–2), 71–120.

9-Keele, L. (2008). *Semiparametric regression for the social sciences*. John Wiley & Sons.

10-Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

11-Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

12-Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC.