Proceedings of the 23rd International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017

Nonlinear Dynamic Predictive Model Selection and Interference Using Information Criteria

Yuanlin Gu, Hua-Liang Wei*, Michael A. Balikhin
Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield, United Kingdom
ygu11@sheffield.ac.uk; w.hualiang@sheffield.ac.uk (*corresponding author)

Abstract— **Model selection is a crucial step in choosing a best model from a series of candidate models for data based modelling problems. The commonly used Akaike information criterion (AIC) and Bayesian information criteria (BIC) may not be effective for many real-world modelling problems when the true system model structure is unknown and therefore not included in the candidate model set. This study investigates the model selection issue using AIC, BIC and an adjustable prediction error sum of squares (APRESS) for nonlinear dynamic predictive modelling. Results from simulation and real data modelling case studies show that both BIC and APRESS produce good models for nonlinear modelling problems. The APRESS works slightly better than BIC in achieving a parsimonious representation of the studied system. In addition, a model averaging method is introduced, which is capable to provide an averaged model that is more robust in generalization (i.e. in representing future data) than any single model.**

Keywords-model selection; model averaging; nonlinear modelling

I. INTRODUCTION

Model selection is a crucial step in data based modelling problems. Given a set of data, main analysis objective is often to build a number of candidate models and determine which of the models best approximates the data. For both linear and nonlinear modelling problems, the model selection criterion is usually required to be able to select a model that 1) fits the data well, 2) consists of model variables that can be easily interpreted, 3) involves a parsimonious representation, and 4) can be used for inference and model prediction [1].

Of many methods, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are the most commonly used measures for model selection. AIC was first proposed in 1974 [2] and designed to approximately estimate the Kullback-Leiber information of a model [3]. AIC is calculated for every candidate model and the model with the smallest AIC is possibly the best choice to approximate the data. Later in 1989, the second-order Akaike information criterion (AIC_c) was developed for small sample size data modelling problems [4][5]. For some model selection problems, it is necessary to measure how much better the best model is when compared with the other models. Typically, two AIC measures can be used: the delta AIC and the Akaike weights. The delta AIC simply calculates the difference between the smallest AIC value and AIC values of other candidate models [6]. The AIC weight is a value between 0 and 1, and can be considered as analogous to the probability that a candidate model is the best choice [7]. Over the past few decades, AIC has increasingly been used for model selection in a wide range of fields, for example, ecology [8], phylogenetics [9], etc. Some model averaging techniques were also developed based on AIC framework, for example, the natural averaging method [7] and full model averaging method [10]. Bayesian information criterion (BIC), proposed by Schwarz [11], also referred to as the Schwarz information criterion, or the Schwarz Bayesian information criterion, is another widely used information criterion for model selection. Similar to AIC, the value of BIC also needs to be calculated for each of candidate models and the minimum of BIC values suggests the best model [12]. There are also plenty of studies that use BIC as model selection criterion for many applications [13][14].

In fact, AIC and BIC are two very similar information criteria. The only difference is that BIC uses a greater penalty than AIC, on the increment of model terms. Thus, a question raises: which of the two model selection criteria is better? Although there are plenty of comparative studies supporting both AIC and BIC (see for examples [15]-[17]), it is not sensible to simply conclude that one is better than another, regardless of the application scenarios. Based on the large literature of applications of AIC and BIC [8][9][13][14], it is reasonable to believe that the data, model type and other aspects of the modelling problems can be significantly important determining which of the criteria is more suitable. In fact, previous studies of using AIC and BIC for model selection usually focused on linear modelling problems or nonlinear modelling problems when the model structure or some prior information is known. With the assumption that the 'true' model is among the candidate models, AIC and BIC can usually give good and consistent indications for the correct model. However, it still lacks evidence that the two criteria also work well for complex system identification problems when the "true" system model can never be known and any candidate models can only give an approximation to the true system model in some degree. In these situations, either AIC or BIC may fail to select the best model from a specified candidate model set. Based on the these considerations, it is essential to investigate the model selection issue using AIC and BIC for general nonlinear system identification problems.

For many nonlinear system identification and data based modelling problems, an effective model selection approach is the cross-validation (CV) based criterion [18].

Two commonly used variations of CV are the Leave-One-Out (LOO), also called Predicted Sum of Squares (PRESS) [19][36][37], and generalised cross-validation (GCV) [20]. A modified generalised cross-validation criterion, also known as adjusted predicted sum of squares (APRESS), was proposed in for nonlinear systems identification [21]. The main advantage of APRESS is that it can be easily implemented and can produce good model selection indications in many applications [21][31].

This study first investigates the performances of three model selection criteria (AIC, BIC and APRESS) for nonlinear dynamic model identification, and then introduces a model averaging approach that help improve the model robustness. A simulation example and a real data example are used to illustrate the performance of the model selection and model averaging methods. The candidate models in the two case studies are chosen to be the nonlinear autoregressive moving average with exogenous input model (NARMAX) [23] structure and are estimated by an orthogonal search algorithm [22].

The reminder of this paper is organised as follows. In Section 2, model selection criteria (AIC, BIC and APRESS) are briefly reviewed and applied for model selection tasks of the two given examples. In Section 3, model averaging approaches are introduced to improve the model robustness and performances. The paper is concluded in Section 4.

II. Evaluation of Model Selection Performance for Nonlinear Dynamic Model Identification

This section first briefly reviews three model selection criteria (AIC, BIC and APREE), and two case studies are carried out to evaluate the performances of these model selection criteria.

A. Model selection with AIC and BIC

The calculation of AIC and BIC is straightforward. AIC is calculated using the number of fitted parameters in the model (designated by k) and the maximum likelihood estimate for the model (designated by L), whereas BIC uses an extra measure which is the sample size n. AIC and BIC can be calculated as [2][11]:

$$\mathrm{AIC} = -2\ln(\mathrm{L}) + 2\mathrm{k} \tag{1}$$

$$\mathrm{BIC} = -2\ln(\mathrm{L}) + \mathrm{k}\ln(\mathrm{n}) \tag{2}$$

Both criteria contain two components: the first one, $-2\ln(\mathrm{L})$, is the value of the likelihood function, indicating the probability of obtaining the data given the model. The second component is used to penalize the model when more model terms (some time called parameters in statistics) are added to the model. Therefore, there is a tradeoff between the better fit and the model complexity. It can be noted that the only difference between AIC and BIC is that BIC uses a greater penalty on the increment of the model parameters. Comparing the AIC and BIC values of all the candidate models, the model with the minimum of AIC or BIC is possibly the 'best' model. For least square based regression analysis, AIC and BIC can be calculated by using the residual sum of squares (RSS), as [4]:

$$\mathrm{AIC} = \mathrm{n}\ln\left(\frac{\mathrm{RSS}}{\mathrm{n}}\right) + 2\mathrm{k} \tag{3}$$

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

For small sample size problem, the second-order AIC ($\mathrm{AIC_c}$) is recommended. The $\mathrm{AIC_c}$ is defined as [4]:

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2\mathrm{k}+1}{\mathrm{n}-\mathrm{k}-1} \tag{5}$$

For large dataset, $\mathrm{AIC_c}$ is very close to AIC. Thus, it is often considered for data modelling problems of small sample size problems. In regular, both AIC and BIC decrease when a first few model terms are included in the model, as a result of the reduction of prediction error. After an enough number of model terms are added, the penalty component becomes significant, and therefore the values of both criteria start to increase. Thus, the model with a minimum AIC or BIC value is then treated to be an optimal choice with both good prediction performance as well as parsimonious representation of the system.

B. NARMX model and APRESS criterion

The adjustable prediction error sum of squares (APRESS) was initially introduced and incorporated to a forward orthogonal search procedure [21]. The combined method is called the adaptive orthogonal search (AOS) and can be used to select significant model terms according to an error reduction ratio index (ERR), and estimate model parameters simultaneously [22][38]. The orthogonal search algorithm, also called orthogonal least squares (OLS) algorithm, has been applied to many research scenarios for term selection and parameter estimation.

A wide range of nonlinear systems can be well represented using the nonlinear autoregressive moving average with exogenous model (NARMAX) model structure [23], which can be described as [23]:

$$\mathrm{y}(\mathrm{t}) = \mathrm{F}\big[\mathrm{y}(\mathrm{k}-1), \dots, \mathrm{y}(\mathrm{k}-\mathrm{n_y}), \mathrm{u}(\mathrm{k}-1), \dots, \mathrm{u}(\mathrm{k}-\mathrm{n_u}), \mathrm{e}(\mathrm{k}-1), \dots, \mathrm{e}(\mathrm{k}-\mathrm{n_e})\big] \tag{8}$$

where $\mathrm{y}(\mathrm{k})$ and $\mathrm{u}(\mathrm{k})$ are systems output and input signals; $\mathrm{e}(\mathrm{k})$ is a noise sequence which is with zero-mean and finite variance. $\mathrm{n_y}$, $\mathrm{n_u}$ and $\mathrm{n_e}$ are the maximum lags for the system output, input and noise. $\mathrm{F}[\cdot]$ is some nonlinear function. A polynomial NARX model can be written as the following linear-in-the-parameters form:

$$\mathrm{y}(\mathrm{k}) = \sum_{\mathrm{m}=1}^{\mathrm{M}} \theta_\mathrm{m}\varphi_\mathrm{m}(\mathrm{k}) + \mathrm{e}(\mathrm{k}) \tag{9}$$

where $\varphi_\mathrm{m}(\mathrm{k}) = \varphi_\mathrm{m}(\vartheta(\mathrm{k}))$ are the model terms generated from the regressor vector $\vartheta(\mathrm{k}) = [\mathrm{y}(\mathrm{k}-1), \dots, \mathrm{y}(\mathrm{k}-\mathrm{n_y}), \mathrm{u}(\mathrm{k}-1), \dots, \mathrm{u}(\mathrm{k}-\mathrm{n_u})]^{\mathrm{T}}$, θ_m are the unknown paramters and M is the number of candidate model terms. The NARMAX model and the OLS algorithm have been applied to successfully solve a wide range of real world problems in various fields including ecological [25], environmental [26], geophysical [24][28][29], medical [30], societal [31] and neurophysiological [27] sciences. The APRESS can be easily calculated in each term selection step in OLS algorithm. It is defined as [21]:

$$\mathrm{APRESS}(\mathrm{k}) = \mathrm{p}(\mathrm{k})\mathrm{MSE}(\mathrm{k}) = \left(\frac{1}{1-\frac{\mathrm{C}(\mathrm{k},\alpha)}{\mathrm{N}}}\right)^2 \mathrm{MSE}(\mathrm{k}) \tag{10}$$

where $\mathrm{MSE}(\mathrm{k})$ is the mean square error of the candidate model, $\mathrm{C}(\mathrm{k},\alpha) = \mathrm{k}\times\alpha$ is a cost function of model

complexity and p(k) is the penalty function. The details about the calculation of APRESS in OLS algorithm can be found in [21][35].

C. A simulation example

Consider a nonlinear system described by the model below:

$$y(t) = -u(t-1)\sqrt{|y(t-1)|} + 0.5u^2(t-1) + u^2(t-2) + y(t-2)u(t-1) + \xi(t) \quad (11)$$

where the input $u(t)$ was assumed to be uniformly distributed on $[-1, 1]$, and the noise $\xi(t)$ was determined by:

$$\xi(t) = w(t) + 0.3w(t-1) + 0.6w(t-2) \quad (12)$$

with $w(t)$ being chosen to be a Gaussian white noise $w(t) \sim N(0, 0.01^2)$. A total number of 1000 input-output data points were generated. The first 500 points were used for model estimation and selection and the second 500 points were used for performance test. A regression vector can be defined as:

$$\varphi(t) = [y(t-1), y(t-2), u(t-1), u(t-2)]^T \quad (13)$$

with the maximum time lags of $n_y = n_u = 2$. The initial full model was chosen to be a polynomial form described by (9) with nonlinear degree of $l = 2$. There are totally 20 candidate model terms, including the constant term. It can be noted that the model term $\sqrt{|y(t-1)|}$ was not in the candidate terms. The orthogonal search algorithm was used to select model terms and estimate the model, and the AIC, BIC and APRESS were used to evaluate all the candidate models. The AIC, BIC and APRESS statistics with different model lengths were calculated and shown in Fig 1.

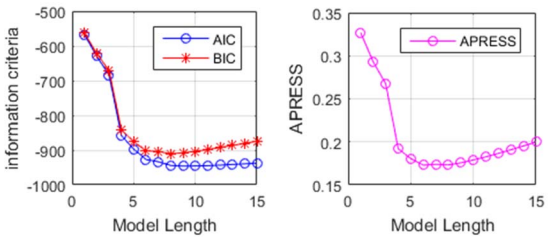

Figure 1. AIC, BIC and APRESS statistics with different model length

Some statistical evaluations of the models suggested by AIC, BIC and APRESS are shown in Table I, where the 'best' models suggested by the smallest AIC, BIC and APRESS were represented by $\text{Model}_{\text{AIC}}$, $\text{Model}_{\text{BIC}}$ and $\text{Model}_{\text{APRESS}}$, respectively. The three models were evaluated using the correlation coefficient (r), predict efficiency (PE) and normalized root mean square error (NRMSE) of model predicted output (MPO). The notation 'length' represents the number of the terms included in each model.

It can be seen that the number of model terms suggested by AIC and BIC is much larger than that in the true model. Compared with AIC and BIC, the APRESS suggests a better choice of 6 model terms, which is smaller than that suggested by AIC and BIC. Also, the model suggested by

TABLE I. EVALUATION OF MODELS SELECTED BY AIC, BIC AND APRESS

Model	length	r	PE	NRMSE
Model_AIC	9	0.6050	0.3645	0.1157
Model_BIC	8	0.6056	0.3654	0.1162
Model_APRESS	6	0.6303	0.3931	0.1130

APRESS, although with fewer number of model terms, possesses slightly better predicative capability. It should be noted that the prediction performances can be affected by the uncertainty caused by noises. Thus, it is normal that any of the models can achieve slightly better statistics of correlation, prediction efficiency and error, as long as they includes the main components of the true model. However, it is also crucially important to achieve a parsimonious representation for complex nonlinear systems in many application situations, due to the fact that a model with less variables can largely reduce the work of data collection and benefit the process of understanding the systems. In general, all the three model selection criteria are capable for model selection for this example. It is possibly because that although the model term $\sqrt{|y(t-1)|}$ is not in the candidate term set, it can be derived from the model term $y(t-1)$. Note that in consideration of model parsimony, APRESS is a better choice for nonlinear dynamic predictive model selection.

D. A real world application: Dst index forecast

The magnetosphere can be considered as a complex system. In order to understand the magnetosphere system, a Dst index is often used to measure the magnetic disturbances [32] [33]. In this study, the process of Dst is treated to be an unknown nonlinear system, where the system inputs are solar wind variables and the system output is the Dst index. The description of the inputs and output is given in Table II. All the variables were sampled every 1 hour. It should be noted that VBS is a multiplied input which was suggested to be included in the model inputs [34].

TABLE II. DST INDEX AND SOLAR WIND VARIABLES

Name	Description
Dst	Dst index
Bs	magnitude of the interplanetary magnetic field
VBS	$V \times Bs/1000$
p	solar wind pressure (flow pressure) [nPa]
V	solar wind speed/velocity (flow speed) [km/s]

The data of Jun 1998 was used to train the model and the data of Jul 1998 was used for model evaluation. Similar to the previous discussed simulation example, the orthogonal search algorithm was used to select model terms and estimate the model parameters, and the AIC, BIC and APRESS were used for model selection. The time lag of inputs was chosen to be 4 and the nonlinear degree was 2. Note that the model is input-alone model (Volterra model), meaning that no autoregressive model terms were included in the inputs. There were totally 16 candidate model terms and 50 candidate models were estimated. All the models are

used to predict Dst index 1 hour ahead. The AIC, BIC and APRESS statistics with different model lengths are shown in Fig 2.

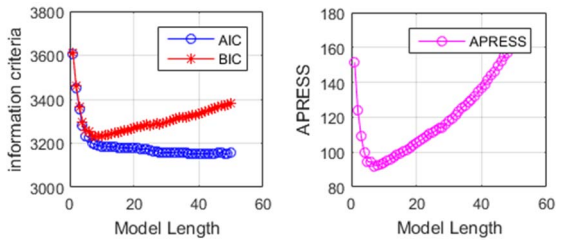

Figure 2. AIC, BIC and APRESS statistics with different model lengths

The number of model terms suggested by AIC, BIC and APRESS are 38, 9 and 7, respectively. The evaluation of the prediction performances of the three models are shown in Table III.

TABLE III. EVALUATION OF MODELS SELECTED BY AIC, BIC AND APRESS

Model	length	r	PE	NRMSE
Model_AIC	38	0.5876	0.2296	0.1362
Model_BIC	9	0.7571	0.5491	0.1041
Model_APRESS	7	0.7521	0.5405	0.1052

It is clear that AIC fails to select the 'best' candidate model. The model with 38 terms performs poorly in forecasting Dst index 1 hour ahead. On the contrary, the models chosen by BIC and APRESS are quite similar and achieve very similar performances. Additionally, the model suggested by APRESS involves a relatively smaller number of model terms. Clearly, for this real data example, both BIC and APRESS are capable for the model selection task. If a parsimonious representation is required, the APRESS statistic is superior to the other two model selection criteria.

III. EVALUATION OF MODEL AVERAGING PERFORMANCE FOR NONLINEAR DYNAMIC MODEL IDENTIFICATION

This section first briefly introduces model averaging approaches based on the three model selection criteria and then produces averaged models for the simulation and real data examples in section II.

A. Model averaging with AIC, BIC and APRESS

The model averaging approach with AIC involves the computation of the delta AIC and the Akaike weights. The delta AIC and the Akaike weight are used to assess the strength of evidence for each candidate model. The delta AIC can be calculated as [6]:

$$\Delta AIC_{c_i} = AIC_{c_i} - AIC_{c_{min}} \tag{14}$$

where AIC_{c_i} is the AIC value for the i^{th} candidate model, $\mathrm{AIC}_{c_{min}}$ is the minimum AIC of all the M candidate models, and $i = 1, 2, \dots, M$. The Akaike weight indicates the probability that an individual candidate model is the best model. The Akaike weight for i^{th} candidate mode is computed as [7]:

$$\omega_i = \frac{\exp(-0.5\Delta \mathrm{AIC}_{c_i})}{\sum_{j=1}^{M} \exp(-0.5\Delta \mathrm{AIC}_{c_j})} \tag{15}$$

where ω_i is the Akaike weight for the i^{th} candidate model and $i = 1, 2, \dots, M$. Then, the averaged parameter estimate is calculated as follows:

$$\hat{\bar{\beta}} = \frac{\sum_{i=1}^{R} \omega_i \hat{\beta}_i}{\sum_{i=1}^{R} \hat{\beta}_i} \tag{16}$$

where $\hat{\beta}_i$ is the estimate of parameter of the i^{th} candidate model and R is the number of candidate models that the averaging inference is based on ($R < M$). This approach is called 'natural averaging'. However, there is some issue when the model selection uncertainty is large. Thus, the model averaging inference needs to be based on all the candidate models. The above formula can simply be reduced to:

$$\hat{\bar{\beta}} = \sum_{i=1}^{M} \omega_i \hat{\beta}_i \tag{17}$$

The second approach is referred to as 'full model averaging'. Here, all the candidate models give their contributions to the averaged parameters according to their Akaike weights. Note, in contrast to the single 'best' model selected by the smallest AIC, the averaged model is often more robust to the peak values of frequency in the data. This is because a single model only uses a limit number of model terms suggested by model selection criterion, but does not consider the effect of other model terms. These model terms may not be significant for the most of the sample points, but can be crucially important for some extreme values. To produce averaged model based on BIC and APRESS weights, a simple approach is to use the full model averaging method. The AIC values can be replaced by BIC and APRESS, to calculate the BIC and APRESS weights of model parameters of all candidate models. The averaged parameters can then be computed using formula (17). This method is simple to implement. More importantly, it is easy to observe which of the three criteria gives the best averaged model. .

B. The simulation data

The time lags and nonlinear degree remain the same as in section II B. A total number of 15 candidate models were estimated using the orthogonal search algorithm. The averaged parameters were calculated based on all the 15 candidate models using formula (17). Note that all the three averaged models were calculated from the same 15 candidate models. The only difference is that the averaged parameter was computed using different weights based on AIC, BIC and APRESS, respectively.

A comparison of the performances of the three averaged models is shown in Table IV. In the table, the averaged models computed based of AIC weights, BIC weights and APRESS weights are represented by $\text{Averaged Model}_{\text{AIC}}$, $\text{Averaged Model}_{\text{BIC}}$ and $\text{Averaged Model}_{\text{APRESS}}$, respectively.

Comparing Table I and Table IV, it can be observed that performance of the three single models and the associated averaged models are similar. The averaged models of AIC

and BIC are slightly better than the associated single models but the averaged model of APRESS is not as good as the single model. The reason is that the single model

TABLE IV. EVALUATION OF AVERAGED MODELS BY AIC, BIC AND APRESS

Model	r	PE	NRSME
Averaged $\text{Model}_{\text{AIC}}$	0.6122	0.3735	0.1149
Averaged $\text{Model}_{\text{BIC}}$	0.6066	0.3668	0.1159
Averaged $\text{Model}_{\text{APRESS}}$	0.6136	0.3668	0.1149

selected by APRESS has better performance than the models selected by AIC and BIC. But the averaged models use information of all the candidate models. In the process of computing the averaged model based on APRESS weights, the information of a few more candidate models were considered and the performances of these models are worse than the single model selected by APRESS, thus the performance of averaged model is decreased. On the contrary, the 'better' single model (single model selected by APRESS) was included in the process of computing averaged models based on AIC and BIC weights. As result of the parameter averaging, the performance of averaged models were improved.

However, the averaged model of APRESS is still the best model among the three averaged models, due to the fact that the best single model has the largest weights in APRESS model averaging process. Based on the above discussion, it can be inferred that the model averaging method can improve the model robustness to avoid the risk that a single model suggested by a model selection method may not be a good choice to represents the data. Therefore, it is strongly suggested that model averaging methods are applied for nonlinear dynamic predictive modelling problems.

C. A real world application: Dst index forecast

For the Dst forecast problem, a total number of 50 candidate models were estimated by the orthogonal search algorithm. The averaged parameters were calculated for the candidate models based on AIC, BIC and APRESS weights. The result of the three averaged models is shown in Table V.

TABLE V. EVALUATION OF AVERAGED MODELS BY AIC, BIC AND APRESS

Model	r	PE	NRSME
Averaged $\text{Model}_{\text{AIC}}$	0.5920	0.2386	0.1354
Averaged $\text{Model}_{\text{BIC}}$	0.7546	0.5460	0.1046
Averaged $\text{Model}_{\text{APRESS}}$	0.7573	0.5460	0.1037

It can be seen that the performances of the averaged models are also similar to the associated single models. Following the discussion above, it can be concluded that the model averaging approaches is consistent with the model selection results for these two examples. The performance of the averaged model is mainly affected by the 'best' single model chosen by AIC, BIC or APRESS, while the other candidate models make smaller contribution to the averaged model according to the relevant averaged models. When the model selection approaches fail to select the best model, the model averaging method can usually help to improve the model performance.

IV. CONCLUSION

This study investigates the model selection issues for system identification and dynamic predictive nonlinear modelling problems. Three model selection criteria (AIC, BIC and APRESS) were evaluated using a simulation example and a case study on real data. Comparing the performances of the models suggested by the three criteria, it can be concluded that BIC and APRESS work well for both the simulation and real data modelling problems. In consideration of the model parsimony, APRESS works better in reducing the number of model terms and can achieve a satisfied prediction performance. Another advantage of APRESS is that it is simple to compute directly in orthogonal search procedure. In conclusion, this study suggests using APRESS for model selection in nonlinear system identification, especially for real data based modelling problems when the true system model structure is unknown and therefore is not in the set of candidate models. In addition, a model averaging approach based on full averaging method is also advised for producing robust averaged models.

ACKNOWLEDGMENT

The authors acknowledge that this work was supported in part by EU Horizon 2020 Research and Innovation Programme Action Framework under grant agreement 637302, the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/I011056/1 and Platform Grant EP/H00453X/1.

REFERENCES

[1] Preacher, K. J. and Merkle, E. C., The problem of model selection uncertainty in structural equation modeling, Psychological methods, 2012, 17(1), 1.

[2] H. Akaike, A new look at the statistical model identification, IEEE Transactions on automatic control, 1974, 19(6), 716-723.

[3] H. Akaike, Information theory and an extension of the maximum likelihood principle, Selected Papers of Hirotugu Akaike, Springer New York, 1998, 199-213

[4] C. M. Hurvich and C. L. Tsai, Regression and time series model selection in small samples, Biometrika, 1989, 297-307.

[5] P. J. Brockwell and R. A. Davis, Time series: theory and methods, Springer Science & Business Media, 2013.

[6] M. R. Symonds and A. Moussalli, A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike information criterion, Behavioral Ecology and Sociobiology, 2011, 65(1), 13-21.

[7] S. T. Buckland, K. P. Burnham and N. H. Augustin, Model selection: an integral part of inference, Biometrics, 1997, 603-618.

[8] J. B. Johnson and K. S. Omland, Model selection in ecology and evolution, Trends in ecology & evolution, 2004, 19(2), 101-108.

[9] D. Posada and T. R. Buckley, Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over

likelihood ratio tests, Systematic biology, 2004, 53(5), 793-808.

[10] P. M. Lukacs, K. P. Burnham and D. R. Anderson, Model selection bias and Freedman's paradox, Annals of the Institute of Statistical Mathematics, 2010, 62(1), 117-125.

[11] G. Schwarz, Estimating the dimension of a model, The annals of statistics, 1978, 6(2), 461-464.

[12] R. E. Kass and A. E. Raftery, Bayes factors, Journal of the american statistical association, 1995, 90(430), 773-795.

[13] M. B. Hooten and N. T. Hobbs, A guide to Bayesian model selection for ecologists, Ecological Monographs, 2015, 85(1), 3-28.

[14] H. L. Wei, S. A. Billings and M. A. Balikhin, Wavelet based non-parametric NARX models for nonlinear input output system identification, International Journal of Systems Science, 2006, 37(15), 1089-1096.

[15] C. A. Medel and S. C. Salgado, Does the BIC Estimate and Forecast Better than the AIC?, Revista de Análisis Económico–Economic Analysis Review, 2013, 28(1), 47-64.

[16] K. Aho, D. Derryberry and T. Peterson, Model selection for ecologists: the worldviews of AIC and BIC, Ecology, 2014, 95(3), 631-636.

[17] 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, 2012, 17(2), 228.

[18] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the royal statistical society. Series B (Methodological), 1974, 111-147.

[19] D. M. Allen, The relationship between variable selection and data agumentation and a method for prediction, Technometrics, 1974, 16(1), 125-127.

[20] G. H. Golub, M. Heath and G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter, Technometrics, 1979, 21(2), 215-223.

[21] S. A. Billings and H. L. Wei, An adaptive orthogonal search algorithm for model subset selection and non-linear system identification, International Journal of Control, 2008, 81(5), 7.

[22] S. Chen, S. A. Billings and W. Luo, Orthogonal least squares methods and their application to non-linear system identification. International Journal of Control, 1989, 50(5), 1873-1896.

[23] S. Chen and S. A. Billings, Representations of non-linear systems: the NARMAX model, International Journal of Control, 1989, 49(3), 1013-1032.

[24] M. A. Balikhin, R. J. Boynton, S. N. Walker, J. E. Borovsky, S. A. Billings, H. L. Wei, Using the NARMAX approach to model the evolution of energetic electrons fluxes at geostationary orbit, Geophysical Research Letters, 2011, 38 (18).

[25] A. M. Marshall, G. R. Bigg, S. M. van Leeuwen, J. K. Pinnegar, H. L. Wei, T. J. Webb, and J. L. Blanchard, Quantifying heterogeneous responses of fish community size structure using novel combined statistical techniques, Global Change Biology, 2015.

[26] G. R. Bigg, H. L. Wei, D. J. Wilton, Y. Zhao, S. A. Billings, E. Hanna, and V. Kadirkamanathan, A century of variation in the dependence of Greenland iceberg calving on ice sheet surface mass balance and regional climate change, Proc. R. Soc. A, 2014, 470(2166), 20130662.

[27] Y. Li, H. L. Wei, S. A. Billings, P. G. Sarrigiannis, Identification of nonlinear time-varying systems using an online sliding-window and common model structure selection (CMSS) approach with applications to EEG, International Journal of Systems Science, 2016, 47(11), 2671-2681.

[28] H. L. Wei, S. A. Billings, A. S. Sharma, S. Wing, R. J. Boynton, and S. N. Walker, Forecasting relativistic electron flux using dynamic multiple regression models, Annales Geophysicae, 2011, 29(2), 415-420.

[29] R. J. Boynton, M. A. Balikhin, S. A. Billings, H. L. Wei, N. Ganushkina, Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetospher, Journal of Geophysical Research: Space Physics, 2011, 116 (A5).

[30] C. G. Billings, H. L. Wei, P. Thomas, S. J. Linnane, and B. D. M. Hope-Gill, The prediction of in-flight hypoxaemia using non-linear equations, Respiratory medicine, 2013, 107(6),841-847.

[31] Y. Gu and H. L. Wei, Analysis of the relationship between lifestyle and life satisfaction using transparent and nonlinear parametric models, 22nd International IEEE Conference on Automation and Computing, 2016, 54-59.

[32] H. L. Wei, S. A. Billings and M. Balikhin, M, Prediction of the Dst index using multiresolution wavelet models, Journal of Geophysical Research: Space Physics, 2004, 109(A7).

[33] H. L. Wei, D. Q. Zhu, S. A. Billings and M. A. Balikhin, Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks, Advances in Space Research, 2007, 40(12), 1863-1870.

[34] W. D. Gonzalez, J. A. Joselyn, Y. Kamide, H. W. Kroehl, G. Rostoker, B. T. Tsurutani, and V. M. Vasiliunas, What is a geomagnetic storm?, J. Geophys. Res., 1994, 99(A4), 5771–5792.

[35] H. L. Wei and S. A. Billings, Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information, International Journal of Modelling, Identification and Control, 2008, 3(4), 341-356.

[36] X. Hong, P. M. Sharkey and K. Warwick, A robust nonlinear identification algorithm using PRESS statistic and forward regression, IEEE Trans. Neural Networks, 2003, 14, 454–458.

[37] S. Chen, X. Hong, C. J. Harris and P. M. Sharkey, Sparse modeling using orthogonal regression with PRESS statistic and regularization, IEEE Trans. Syst. Man, Cyber. B, 2004, 34, 898–911.

[38] H. L. Wei, S. A. Billings and J. Liu, Term and variable selection for non-linear system identification, International Journal of Control, 2004, 77 (1), 86-110.