Order Selection for General Expression of Nonlinear Autoregressive Model Based on Multivariate Stepwise Regression

Jinfei Shi, Songqing Zhu, Ruwen Chen
Nanjing Institute of Technology, Nanjing, China, 211167

Abstract. An order selection method based on multiple stepwise regressions is proposed for General Expression of Nonlinear Autoregressive model which converts the model order problem into the variable selection of multiple linear regression equation. The partial autocorrelation function is adopted to define the linear term in GNAR model. The result is set as the initial model, and then the nonlinear terms are introduced gradually. Statistics are chosen to study the improvements of both the new introduced and originally existed variables for the model characteristics, which are adopted to determine the model variables to retain or eliminate. So the optimal model is obtained through data fitting effect measurement or significance test. The simulation and classic time-series data experiment results show that the method proposed is simple, reliable and can be applied to practical engineering.

1. Introduction
Time series analysis technology is a kind of system identification method which can establish models based on the inherent law of data with no need for system inputs, so it has important applications in natural and social science field of industrial process control, economy and biomedical engineering, etc.[1]-[5].
Determining model's order plays a very important role in time series analysis process. Insufficient order of the model tends to the losing of valuable information to decrease model's tracking and prediction ability of data sequence. Meanwhile, excessive order makes a complex model and huge computing quantity.
At present, order determination for the linear models based on the stationary assumptions have many research and lots of classic criterions are applied in engineering such as AIC criterion, residual test, minimization of residual sum of squares or final prediction error and other information criterions[6][7]. But for nonlinear systems, due to their multifarious characteristics, there are no universal order determinations or goodness judgment strategies.
The General Expression of Nonlinear Autoregressive (GNAR) model is a novel time-series model, which can be applied to system identification, data tracking and system prediction, etc.[8][9]. In this paper, an order selection method based on multiple stepwise regressions is proposed for GNAR model which converts the model order selection into variable selection for multiple linear regression equation. Firstly, the partial autocorrelation function is adopted to define the linear term in GNAR model. And the linear result is set as the initial model, and then the nonlinear terms are added gradually. Statistics are chosen to study all the variables and their improvements are adopted to determine the model variables to retain or eliminate. Finally, the optimal model is obtained through data fitting effect measurement or significance test.
This remainder paper is organized as follows. Section 2 introduces the definition of GNAR model, Section 3 discusses the order selection steps based on multivariate stepwise regression for linear and nonlinear terms in GNAR model, Section 4 describes its application and finally Section 5 presents the conclusion.

2. Definition of GNAR model
The expression of GNAR model can be expressed as:

\[
w_t = \sum_{i=1}^{\infty} \alpha_{i} w_{t-i} + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_{i,j} w_{t-i} w_{t-j} + \cdots + \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \cdots \sum_{l=1}^{\infty} \alpha_{i,j,k, \cdots, l} \prod_{r=1}^{l} w_{t-r} + \epsilon_t
\]

where \( w_{t,i} \) represents the system observed data at the moment of \( t-i, i=0, 1, 2, \ldots \); \( \alpha_{i}, \alpha_{i,j}, \alpha_{i,j,k, \cdots, l}, \cdots \) are model parameters; \( \epsilon_t \) is white noise with zero mean.

When modeling in engineering, the model order should not be infinite, so Eq.(1) is rewritten as:

\[
w_t = \sum_{i=1}^{n_1} \alpha_{i} w_{t-i} + \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \alpha_{i,j} w_{t-i} w_{t-j} + \cdots + \sum_{i=1}^{n_p} \sum_{j=1}^{n_p} \sum_{k=1}^{n_p} \cdots \sum_{l=1}^{n_p} \alpha_{i,j,k, \cdots, l} \prod_{r=1}^{l} w_{t-r} + \epsilon_t
\]

in which \( p \) is the polynomial order; \( \sum_{i=1}^{n_1} \alpha_{i} \) is the first order linear term, \( \alpha_{i} \ (i_1=1,2, \ldots, n_1) \) are linear coefficients; \( \sum \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \alpha_{i,j} w_{t-i} w_{t-j} \) is the second order nonlinear term, \( \alpha_{i,j} \ (i_2, j_2=1,2, \ldots, n_2) \) are the second order nonlinear coefficients; \( \sum \sum \sum \cdots \sum_{i=1}^{n_p} \sum_{j=1}^{n_p} \sum_{k=1}^{n_p} \cdots \sum_{l=1}^{n_p} \alpha_{i,j,k, \cdots, l} \prod_{r=1}^{l} w_{t-r} \) is the \( p \) order nonlinear term, \( \alpha_{i, \cdots, l} \ (i, \ldots, l_p=1,2, \ldots, n_p) \) are the \( p \) order nonlinear coefficients; \( n(j=1,2, \ldots, p) \) are the memory steps of every linear or nonlinear term. The model (2) can be denoted as GNAR \((p; n_1, n_2, \ldots, n_p)\) in abbreviation.

3. Order selection method for GNAR model based on multivariate stepwise regression

3.1. Order selection method
The expansion of Eq. (2) is as follows:

\[
w_t = \alpha_{1} w_{t-1} + \alpha_{2} w_{t-2} + \cdots + \alpha_{n_1} w_{t-n_1} + \alpha_{1,1} w_{t-1}^2 + \alpha_{1,2} w_{t-1} w_{t-2} + \cdots + \alpha_{n_1,n_2} w_{t-n_1}^2 + \alpha_{1,1,1} w_{t-1}^3 + \alpha_{1,1,2} w_{t-1} w_{t-2}^2 + \cdots + \alpha_{n_1,n_2,n_3} w_{t-n_1}^3 + \cdots + \alpha_{1, \cdots, 1, n_p} w_{t-1}^p + \cdots + \alpha_{n_1, \cdots, n_p} w_{t-n_p}^p + \epsilon_t
\]

Let \( W_t = [w_t, w_{t-1}, w_{t-2}, \cdots, w_{t-N+1}, p]^T \), \( \Psi = [\alpha_1, \alpha_2, \alpha_{i_1, \cdots, l_1}, \alpha_{i_2, \cdots, l_2}, \ldots, \alpha_{n_p, \cdots, l_p}]^T \), then GNAR expression for a time series with the length of \( N \) is:
The order selection for GNAR model is to determine the variable matrix in the right of Eq. (4).

\[ W_{t} = \begin{bmatrix} w_{t-1} & \cdots & w_{t-n_{1}} & w_{t-2} & \cdots & w_{t-n_{2}} & \cdots & w_{t-n_{p}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{t-N+p} & \cdots & w_{t-N+p} & w_{t-N+p} & \cdots & w_{t-N+p} & \cdots & w_{t-N+p} \end{bmatrix} \psi + a_{t} \] (4)

Then the order selection for GNAR model is to determine the variable matrix in the right of Eq. (4).

Regard the column vector in this variable matrix as regression variables, and set

\[ X_{t} = [w_{t-1}, w_{t-2}, \ldots, w_{t-n_{p}}]^{T}, X_{t+1} = [w_{t+1}, w_{t+2}, \ldots, w_{t+1}]^{T}, \cdots, X_{N} = [w_{N-1}, w_{N-2}, \ldots, w_{N}]^{T}, \cdots, X_{N-n} = [w_{N-n}, \ldots, w_{N}]^{T}, \cdots, X_{N-n-n} = [w_{N-n-n}, \ldots, w_{N-n-n}]^{T} \]

then:

\[ W_{t} = \begin{bmatrix} X_{1} & X_{2} & \cdots & X_{n_{1}} & \cdots & X_{n_{p-n}} \end{bmatrix} \psi + a_{t} \] (5)

Hence, the order selection for GNAR model is converted to select the variable \( X_{1}, X_{2}, X_{1,1}, \ldots \) in multiple linear regression equation (5). The forward or backward selection or stepwise regression methods based on AIC, BIC, residual mean square, or \( Cp \), can be adopted to resolve variable selection problem [10].

3.2. Order selection for linear term

The coefficients in time series model generally do not have definite physical meaning, but they reflect some inherent characteristics of the system. The coefficients can be considered the weights of each component in the model. Stable system has a relatively stable weight function; especially the coefficients of the linear terms in the model have the robust characteristic [9]. Hence, the memory step of the linear part in GNAR model can be determined by Partial autocorrelation function (PACF) [11].

Take the GNAR(2; 3, 2) shown in Eq. (6) for example. 100 data sequence is generated by simulation, shown in Figure 1.

\[ w_{t} = -0.1w_{t-1} - 0.3w_{t-2} + 0.6w_{t-3} + 0.05w_{t-1}w_{t-1} - 0.04w_{t-1}w_{t-2} - 0.06w_{t-2}w_{t-2} + a_{t} \]

\[ a_{t} \sim N(0, 1) \quad t = 1, 2, \ldots \] (6)

The partial autocorrelation function of this data is obtained, as shown in Figure 2. As can be seen from the Figure 2, within the 95% confidence interval, partial autocorrelation of significance lowers greatly with the delay step of 3. Thus, the rudimentary memory step of the linear part can be set as 3.

3.3. Order selection for nonlinear term
The order and memory steps for nonlinear terms in GNAR model can be selected through stepwise regression method. The rudimentary linear model is set as the initial model and then the nonlinear terms $w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t1}w_{t4}, w_{t1}w_{t5}, \ldots$ are introduced gradually. Statistics are chosen to study both the new introduced and originally existed variables, and their improvements are adopted to determine the model variables to retain or eliminate.

Still take the model of Eq. (6) for example. Least Squares parameter estimation method is adopted. Statistics magnitudes, AIC, BIC and residual mean square value shown in Eq. (7) to Eq. (9), are chosen to make a judgment.

$$AIC = N \ln (SSE / N) + 2p$$

$$BIC = N \ln (SSE / N) + p (\ln N)$$

$$RMS = SSE / (N - p)$$

where, $N$ represents the sample data length; $p$ represents the variable number in regression equation; SSE represents the sum of squared residuals in modeling, and $SSE = \sum (w_i - \hat{w}_i)^2$.

The results of the variable selection for the simulation data are listed in Table 1. It can be learned from the table that GNAR(2; 3, 1) (No 2 in Table 1) is the preferable model according to the statistics of AIC and BIC and GNAR(2; 3, 2) (No 4 in Table 1) to RMS. The models whose AIC values vary between 2 can be regarded as equivalent [2], so model 2, model 3 and model 4 are all suitable to describe this group of data.

| No | Variable | $p$ | SSE   | AIC   | BIC   | RMS   | $R^2$ |
|----|----------|-----|-------|-------|-------|-------|-------|
| 1  | $w_{t1}, w_{t2}, w_{t3}$ | 3   | 129.833 | 32.108 | 39.924 | 1.338 | 0.646 |
| 2  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}$ | 4   | 122.628 | 28.398 | 38.819 | 1.277 | 0.666 |
| 3  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}$ | 5   | 122.221 | 30.066 | 43.092 | 1.287 | 0.667 |
| 4  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t2}w_{t2}$ | 6   | 119.888 | 30.139 | 45.770 | 1.275 | 0.673 |
| 5  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t2}w_{t2}$, $w_{t3}w_{t3}$ | 7   | 119.847 | 32.105 | 50.341 | 1.289 | 0.673 |
| 6  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t2}w_{t2}, w_{t3}w_{t3}, w_{t4}w_{t4}$ | 8   | 119.842 | 34.101 | 54.942 | 1.303 | 0.673 |
| 7  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t2}w_{t2}, w_{t3}w_{t3}, w_{t4}w_{t4}, w_{t5}w_{t5}$ | 9   | 119.842 | 36.100 | 59.547 | 1.317 | 0.673 |
| 8  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t2}w_{t2}, w_{t3}w_{t3}, w_{t4}w_{t4}, w_{t5}w_{t5}, w_{t6}w_{t6}$ | 7   | 119.575 | 31.877 | 50.114 | 1.286 | 0.674 |
| 9  | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t2}w_{t2}, w_{t3}w_{t3}, w_{t4}w_{t4}, w_{t5}w_{t5}, w_{t6}w_{t6}$ | 8   | 119.323 | 33.667 | 54.508 | 1.297 | 0.675 |
| 10 | $w_{t1}, w_{t2}, w_{t3}, w_{t1}w_{t1}, w_{t1}w_{t2}, w_{t1}w_{t3}, w_{t2}w_{t2}, w_{t3}w_{t3}, w_{t4}w_{t4}, w_{t5}w_{5}$ | 9   | 118.806 | 35.232 | 58.679 | 1.306 | 0.676 |

After fitting the given data by GNAR model, we’d better measure model fitting effect for the data or test the model significance further more. Determination coefficient $R^2$ can be used to evaluate the efficacy of the fitting; if the model fits the data well, $R^2$ should close to 1[10].
\[ R^2 = \frac{1 - \text{SSE}}{\text{SST}} \]  

(10)

Determination coefficients of the ten models in Table 1 are shown in Figure 3. \( R^2 \) varies gently after the fourth model which means that GNAR(2; 3, 2) fits the data best and has the less variables. The result is in conformity with the nature of this group of data.

Analysis of variance for model 2, model 3 and model 4 is carried out to test regression effect through F test, and their statistical quantities of F significant are \( 8.206 \times 10^{-22} \), \( 5.188 \times 10^{-21} \), \( 1.402 \times 10^{-20} \) in turn which are far less than the significance level 0.05. So the three model regression effect is remarkable.

From the above, it is feasible to select the GNAR model order based on multiple stepwise regressions. And the process of the order selection method is summarized as shown in Figure 4.
4. Application for classical data

The order selection method proposed is applied to model and analyse the observations sequence of sunspots from 1700 to 1987. According to the steps shown in Figure 4, GNAR model is established based on the 200 sample data from 1700 to 1899. Figure 5 shows the partial autocorrelation function of sunspots sequence from which the linear memory step length can be judged 2.

The Statistic index of variable selection for Sunspots modelling is presented in Table 2. From the table, it can be seen that BIC statistics increases the punishment for model parameters increasing, so the variable selection result stops with GNAR (2, 2, 2); yet the choice of AIC, RMS and $R^2$ is GNAR (2, 2, 7).

Forecasting is the important application of time series analysis technology, so we do the forecasting experiment for sunspot data from 1900 to 1987. Forecasting error square sum named $P_{\text{error}}$ is presented in Table 3. GNAR(2;2,7) has the minimum $P_{\text{error}}$ which confirms that GNAR(2;2,7) is the most suitable model for the group of temporal data.

![Figure 5. The partial autocorrelation function of sunspots sequence.](image)

| No. | Model         | AIC/10^3 | BIC/10^3 | RMS    | $R^2$ |
|-----|---------------|----------|----------|--------|-------|
| 1   | GNAR(1;2)     | 1.244    | 1.251    | 498.132| 0.432 |
| 2   | GNAR(2;2,1)   | 1.236    | 1.246    | 477.153| 0.458 |
| 3   | GNAR(2;2,2)   | 1.221    | 1.237    | 437.706| 0.508 |
| 4   | GNAR(2;2,3)   | 1.218    | 1.245    | 425.935| 0.527 |
| 5   | GNAR(2;2,4)   | 1.217    | 1.257    | 415.347| 0.547 |
| 6   | GNAR(2;2,5)   | 1.211    | 1.267    | 393.328| 0.583 |
| 7   | GNAR(2;2,6)   | 1.207    | 1.283    | 375.529| 0.614 |
| 8   | GNAR(2;2,7)   | 1.198    | 1.297    | 349.092| 0.655 |
| 9   | GNAR(2;2,8)   | 1.232    | 1.357    | 398.915| 0.623 |
| 10  | GNAR(3;2,7,1) | 1.279    | 1.450    | 479.942| 0.587 |
| 11  | GNAR(3;2,7,2) | 1.281    | 1.475    | 475.315| 0.610 |
Table 3. Forecasting error square sum

| No | Model         | P_error/10^4 |
|----|---------------|--------------|
| 1  | GNAR(1;2)     | 4.242        |
| 2  | GNAR(2;2,1)   | 5.087        |
| 3  | GNAR(2;2,2)   | 4.833        |
| 4  | GNAR(2;2,3)   | 4.667        |
| 5  | GNAR(2;2,4)   | 4.385        |
| 6  | GNAR(2;2,5)   | 4.018        |
| 7  | GNAR(2;2,6)   | 3.652        |
| 8  | GNAR(2;2,7)   | 3.372        |
| 9  | GNAR(2;2,8)   | 4.107        |
| 10 | GNAR(3;2,7,1) | 4.552        |
| 11 | GNAR(3;2,7,2) | 4.411        |

5. Conclusion
In this paper, an order selection method based on multiple stepwise regressions is proposed for GNAR model. The process of order selection steps are found out, and the simulation data and the classic time-series data experiment results show that the method proposed is simple, reliable and can be applied to practical engineering.
For a given data set, choosing different statistics perhaps result in different models. Time series data reflect characteristics of the system; therefore, after finding the matching model through some statistics, we should understand the data structure further and analyze the application background to find the optimal model. This is also our future research direction.

Acknowledgments
This work is supported by the National Nature Science Foundation of China (51305194, 51405221), Qing Lan Project (2016) and Jiangsu province science and technology support project (BE2014142). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References
[1] Fan J Q, Yao Q W. Nonlinear Time Series [M]. Beijing: Higher Education Press, 2011.
[2] Yang Shuzai, Wu Ya, Xuan Jianping. Time series analysis in engineering application (the 2nd Edition)[M]. Wuhan: Huazhong University of Science and Technology Press, 2007.
[3] Kalid Yunus, Peiyuan Chen, Torbjörn Thiringer. Modelling spatially and temporally correlated wind speed time series over a large geographical area using VARMA [J]. IET Renewable Power Generation, 2017, 11(1): 132-142.
[4] Pomsak Deekumpa, Itsared Chanjarean, Kissana Audjai. Adaptive time series forecasting model with associate rules from time series application to business and financial[C]. 2017 International Conference on Digital Arts, Media and Technology, 2017, 320-324.
[5] Junpeng Zhan, C. Y. Chung, Elemer Demeter. Time Series Modeling for Dynamic Thermal Rating of Overhead Lines[J]. Power Systems, 2017, 32(3): 2172-2182.
[6] Yao Xiaodong, Huang Yiqiao, Ma Xiaobo, et al. Thermal Error Modeling and Real-Time Compensation of CNC Machine Tools Based on Time Series Method[J]. Journal of Shanghai Jiaotong University, 2016, 5(5): 673-679.
[7] Zhang Liangli, Zhu He, Wu Chaozhong, et al. Risk Estimation to Vehicles Collision at Intersection Based on ARMA Prediction Model [J]. Journal of Transportation Systems Engineering and Information Technology, 2015, 15(5): 239-245.

[8] Chen Ruwen, Wang Yugo, Zhan Shishi. Identification Based on GNAR Model for Vehicle Suspension Anti-vibration Performance [J]. Automotive Technology, 2016, (4): 31-35.

[9] Chen Ruwen. The General Expression for Linear and Nonlinear Auto-regressive Time Series Model and Its Engineering Application [M]. Nanjing: Southeast university press, 2011.

[10] Samprit Chatterjee, Ali S Hadi. Regression Analysis by Example [M]. Beijing: China Machine Press, 2016.

[11] Ruey S Tsay. Analysis of Financial Time Series [M]. Beijing: Posts and Telecom Press, 2012.