International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:3, 2013

GMDH Modeling Based on Polynomial Spline Estimation and Its Applications

LI qiu-min, TIAN yi-xiang, and ZHANG gao-xun

***Abstract*—**GMDH algorithm can well describe the internal structure of objects. In the process of modeling, automatic screening of model structure and variables ensure the convergence rate.This paper studied a new GMDH model based on polynomial spline estimation. The polynomial spline function was used to instead of the transfer function of GMDH to characterize the relationship between the input variables and output variables. It has proved that the algorithm has the optimal convergence rate under some conditions. The empirical results show that the algorithm can well forecast Consumer Price Index (CPI).

***Keywords*—**spline, GMDH, nonparametric, bias, forecast.

I. Introduction

THE Group Method of Data Handling (GMDH) algorithm is a multivariate analysis method for modeling and identifying uncertainty on linear or nonlinearity systems. This algorithm was first introduced by A.G. Ivakhnenko in 1967[1]. The GMDH algorithm uses advantages of both self-organizing principle and multilayer neural networks to select best relationships between variables. First, the GMDH algorithm can automatically find interrelations between variables and chooses the optimal model to fit data. Second, the GMDH algorithm is similar to the neural network, combine with the ideas of black-box, biological neuron method, inductive method, probability theory, and many other methods. The GMDH algorithm unifies automatic control and pattern recognition to reduce human involvement in the process of understanding behavior, and it is objective and impartial [2]-[3].

The main idea of GMDH is the use of feed-forward networks based on short-term polynomial transfer functions whose coefficients are obtained using regression combined with emulation of the self-organizing activity behind NN structural learning (Farlow, 1984). To improve the performance of the GMDH algorithm, Barron (1988) gave a comprehensive overview of some early developments of network, and introduced the Polynomial Network Training algorithm (PNETTR). Elder (1996) proposed Synthesis of Polynomial Network (ASPN) algorithm to improve the GMDH algorithm. J.A.Muller and Frank Lemke (2000) developed and improved self-organizing data mining algorithms on the basis of the above results in 1990s [4]. Further enhancements of the GMDH algorithm have been realized in the "KnowledgeMiner" software. The GMDH algorithm has gradually become an effective tool for modeling, forecasting, and decision support and pattern recognition of complex systems. There are processes for which it is needed to know their future or to analyze inter-relations. The GMDH method has been successfully applied in economy, climate, finance, ecology, medicine, manufacturing and military systems.

Although GMDH provides for a systematic procedure of system modeling and prediction, it also has a number of shortcomings.The traditional GMDH algorithm used Kolmogorov-Gabor polynomial function as the transfer function to create the initial model. When dealing with highly nonlinear systems, it will produce a overly complex network owing to its limited transfer function. Following studies focused on the improvement of the GMDH.

Godfrey C. Onwubolu (2008) using differential evolution in the selection process of the GMDH algorithm, the model building process is free to explore a more complex universe of data permutations [5]. Petr Buryana and Godfrey C. Onwubolu (2011) present an enhanced multilayered iterative algorithm-group method of data handling (MIA-GMDH)-type network [6]. Several specific features such as thresholding schemes and semi-randomised selection approach are used to improving self-organising polynomial GMDH. Tian Y X and Tan D J (2008) used a method of Local Linear Kernel Estimation to improve GMDH modeling for Forecasting [7]. Meysam Shaverdi, Saeed Fallahi, Vahhab Bashiri (2012) presented a GMDH type-neural network based on Genetic algorithm, and used to predict stock price index which are inherently noisy and non-stationary [8].

In this paper we improve GMDH algorithm by incorporating the non-parametric polynomial spline estimation. The proposed non-parametric method does not require any specific assumptions of the relationship between variables, and the results have a good robustness. The polynomial spline function, instead of the Kolmogorov-Gabor polynomial function, is used as the transfer function of GMDH to build up the relationship between input and output variables.

II. The Fundamental of Group Method of Data Handling (GMDH) model

The GMDH was first introduced by Ivakhnenko in the 1960s as a means of identifying nonlinear relations between input and output variables. GMDH has since been applied to a host of

LI qiu-min is with the School of management and Economics, University of Electronic Science and Technology of China, Chengdu610054, China. (corresponding author to provide phone:86-13550061776; e-mail: qiuminlee@hotmail.com).

TIAN yi-xiang is with the School of management and Economics, University of Electronic Science and Technology of China, Chengdu610054, China. (e-mail: tianyx@uestc.edu.cn).

ZHANG gao-xun is with the School of management and Economics, University of Electronic Science and Technology of China, Chengdu610054, China.

practical situations which showed that this class of multilayered polynomial networks has proved effective for both modelling and prediction.

The specific steps involved in the conventional GMDH modeling are:

(1) The sample data set can be divided into the training data set and testing data set.

(2) All possible combinations of the n inputs are generated to create the transfer function $f(X)$ of the $\sum_{l=2}^{n-1} C_n^l$ neurons.The general relationship between input and output variables can be found in the form of a support functional.

$$y = f(x_i, x_j) \qquad (1)$$

The traditional GMDH algorithm used Kolmogorov-Gabor polynomial function as the transfer function to create the initial model. The Kolmogorov-Gabor polynomial function is expressed by:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j\geq i}^{m} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j\geq i}^{m}\sum_{k\geq j}^{m} a_{ijk} x_i x_j x_k + \cdots \qquad (2)$$

where y is the output variables and $X(x_1, x_2, \cdots, x_n)$ is the vector of input variables, $A(a_1, a_2, \cdots, a_n)$ is the vector of the summand coefficients.

(3) The next step is to select an external criterion as the objective function. The GMDH method allows choosing a number of selection criteria, such as the mean root square error criterion. The procedure of inheritance, mutation and selection stop automatically if a new generation of models does not bring any further improvements.

(4) Select $n_1 \leq \sum_{l=2}^{n-1} C_n^l$ variables as new inputs by external criterion and generate all possible combinations of the n_1 inputs to create the transfer function $f(Y)$ of the $\sum_{l=2}^{n_1-1} C_n^l$ neurons of the second layer.

$$z_l = f(y_i, \cdots, y_j) \ , \ i, j = 1,2,\cdots,n_1 \ , \ i \neq j \ ,$$

(5) Repeat the steps 2 to 4. When the errors of the test data in each layer stop decreasing, the iterative computation is terminated.

The aforementioned steps of the GMDH algorithm are executed iteratively until there is no improvement based on the external criterion. The optimal model parameters and model structure will be obtained through pushing back along the last layer.

As mentioned earlier, the traditional GMDH algorithm used Kolmogorov-Gabor polynomial function as the transfer function to create the initial model. A pre-specified relationship between the variables may cause a huge bias and further lead to human error. This paper studies a new GMDH algorithm which is improved by incorporating the polynomial spline estimation. The prediction of the model achieves the desired effect.

III. The Polynomial Spline Estimation

Non-parametric regression method assumes that the relationship between economic variables is unknown; use historical data to estimate the entire regression function [9]. "Spline" comes from the exterior design of the hull and aircraft in engineering. In order to connect the specified sample points into a smooth curve, the spline (i.e. flexible thin strips of wood or thin steel bars) is fixed in the sample points, and then it will be bending freely in other parts. When the curve expressed by spline, called a spline curve or spline function, the sample points called nodes. In mathematics, it is similar to a piecewise cubic polynomial, with first-order and second-order continuous derivative at nodes [10]-[13].

Polynomial spline estimation means that spline function is used to fit the model. The method is a global estimation. It gives a simple explicit expression of the model, and can predict the regression function value of the data outside the region.

Non-parametric model:

$$Y_i = m(X_i) + u_i, \quad i = 1,2,\cdots,n \qquad (3)$$

where X_i is observed value, $m(X_i)$ is an unknown function indicating the complicated underlying relation between inputs and outputs and u_i is the random error.

Suppose $t_1, t_2, \cdots, t_M$ is fixed sequence of nodes. $-\infty < t_1 < t_2 < \cdots < t_M < +\infty$. The basis function of spline function is

$$B_i(x) = (x - t_i)_+^3, (i = 1,2,\cdots,M) \ ,$$
$$B_{M+1}(x) = 1, B_{M+2}(x) = x, B_{M+3}(x) = x^2, B_{M+4}(x) = x^3$$

where $(x - t_i)_+ = \max\{0, x - t_i\}, (i = 1,2,\cdots,M)$,

Polynomial spline function is

$$\sum_{i=1}^{M+4} \beta_i B_i(x) \ , \qquad (4)$$

Minimize

$$\sum_{j=1}^{n}(Y_j - \sum_{i=1}^{M+4}\beta_i B_i(x))^2 \quad (5)$$

Have the estimated value $\hat{\beta}_i (i = 1,2,\cdots,M+4)$ of β_i, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_{M+4})^T$, $Y = (Y_1, Y_2, \cdots, Y_n)^T$,

$$\hat{\beta} = (W^T W)^{-1} W^T Y \quad (6)$$

where

$$W = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \cdots & B_{M+4}(x_1) \\ B_1(x_2) & B_2(x_2) & \cdots & B_{M+4}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n) & B_2(x_n) & \cdots & B_{M+4}(x_n) \end{pmatrix}$$

The polynomial spline estimation of nonparametric regression function $m(X_i)$ is

$$\hat{m}(x) = \sum_{i=1}^{M+4}\hat{\beta}_i B_i(x) \quad (7)$$

In polynomial spline estimation, the choice of nodes is very important. The more the node number, the better the fitting degree of model, the lower the smoothness of the curve. In order to coordinate the trade-off, we should select the appropriate number of nodes. Three common choices for choice of nodes are: Akaike information criterion (AIC), Bayesian information criterion (BIC), and Modified cross-validation criteria (MCV). This paper uses AIC.

$$AIC = \log(RSS/n) + 2*K/n$$

where K is the number of parameters to be estimated, RSS is the residual sum of squares of the formula (5). AIC means that the number of node is automatically selected by minimizing the value of AIC.

IV. GMDH Modeling Based on Polynomial Spline Estimation (SP-GMDH)

This paper uses a non-parametric method to estimate the model instead of pre-specifying a form of the model so as to avoid the possible error during the modeling process. In this paper, the polynomial spline estimation function is used to instead the transfer function of GMDH to build up the relationship between input and output variables. It means that Eq. (3) is used to estimate the model.

The specific steps involved in the k-NN-GMDH model are:

(1) The sample data set W can be divided into the training data set A and testing data set B. y is the output variables and $X(x_1, x_2, \cdots, x_n)$ is the vector of input variables.

(2) In the first layer, the n inputs are generated to all possible combinations $\sum_{l=2}^{n-1} C_n^l$ and are constructed into the transfer function $m(x)$.

$$m(x) = \sum_{i=1}^{M+4}\hat{\beta}_i B_i(x),$$

(3) The screened criterions: Threshold is set to root mean square error (RMSE).

(4)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - o_i)^2} \quad (8)$$

In Equation (8), y_i are forecasted values at data point i, o_i are observed values at data point i. When the value RMSE is smallest, the saved variables as new inputs are constructed the transfer function $m(x)$, continually generate the output variable of the next layer. The process will repeat iteratively until the value RMSE in each layer stop decreasing. The iterative computation is terminated, and the optimal model parameters and model structure will be obtained through pushing back along the last layer.

The performance of the proposed GMDH model based on polynomial spline estimation will be improved in terms of having a better predictive capability than traditional methods.

Assumptions.

A1. The function $m(\cdot)$ has first-order and second-order continuously derivatives.

A2. $K = M + l + 1$, M is the number of the nodes, l is the order of the polynomial, K is the dimension of the polynomial spline space.

A3. G is the polynomial spline function space in compact set. $\rho_{n,j} = \inf_{g \in G} \| g - m \|_2$,

$\rho_n = \max_{i \in \{1,2,\cdots,K\}} \rho_{n,j}$

A4. The eigenvalues of $E(XX^T)$ constant is positive and uniformly bounded.

Lemma 1. Assume that the conditions A1-A4 hold, then

$$\| \hat{m} - m \|_2^2 = O_p(\frac{K}{n} + \rho_n^2),\ j = 1,2,\cdots,n$$

where $K = M + l + 1$, M is the number of the nodes, l is the order of the polynomial.

$\rho_n = \max \rho_{n,j} = \inf \| m_i - m_j \|_2$. In particular, if $\rho_n = o(1)$, then $\hat{m}$ is the consistent estimation of m. That is, $\| \hat{m} - m \|_2 = o_p(1), j = 1,2,\cdots,n$.

Proof [14].

Lemma 2. Assume that the conditions A1-A4 hold, if $m(x)$ has *l*st-order continuously derivatives, and $K = O(n^{1/(2l+1)})$, then

$$\| \hat{m}_j - m_j \|_2 = O_p(n^{-l/(2l+1)}), j = 1,2,\cdots,n \ .$$

Proof [15].

Theorem 1. Assume that the conditions A1-A4 hold, the estimators of polynomial spline function $m(x)$ can achieve the global optimal convergence rate.

Proof. When $l = 2$,

$$\| \hat{m}_j - m_j \|_2 = O_p(n^{-2/5}), j = 1,2,\cdots,n .$$

The global optimal convergence rate is $O_p(n^{-2/5})$ [16]-[18]. That is the polynomial spline estimation can achieve the global optimal convergence rate $O_p(n^{-2/5})$. This rate is faster than the convergence rate $O_p(n^{-4/5})$ in the external point of the Kernel estimation.

When $l = 3$, it is the usual cubic spline function estimation,

$$\| \hat{m}_j - m_j \|_2 = O_p(n^{-3/7}), j = 1,2,\cdots,n .$$

This rate is slower than the convergence rate $O_p(n^{-3/7})$ in the interior point of the Kernel estimation, but faster than the convergence rate $O_p(n^{-4/5})$ in the external point of the Kernel estimation. And this rate maintain globally consistent. Therefore polynomial spline estimation has been proved consistent.

V. An Illustrative Case

The consumer price index (CPI) is the index of price changes in the price level of consumer goods and services purchased by households. The CPI reflects the trend and extent of the consumer price changes in a certain period. The CPI is affected by various society economic factors, including monetary policy, exchange rate, fixed-asset investment, industrial producer prices, agricultural product prices, etc.

TABLE I
CPI, Forecast Results and Relative Error (2010.7-2011.6)

Month	CPI	GMDH forecasting	Relative error (%)	PS-GMDH forecasting	Relative error (%)
July	103.3041	103.2011	-0.0997	103.2896	-0.0140
Aug	103.4794	103.5066	0.0262	103.4937	0.0138
Sept	103.6056	103.6668	0.0591	103.6074	0.0017
Oct	104.3655	103.7729	-0.5678	104.6152	0.2393
Nov	105.1192	104.5364	-0.5544	105.4548	0.3192
Dec	104.5859	105.3122	0.6944	104.6833	0.0931
Jan	104.9000	104.8640	-0.0343	104.9250	0.0238
Feb	104.9443	105.0886	0.1375	105.0149	0.0673
Mar	105.3830	105.4090	0.0247	105.3895	0.0062
Apr	105.3000	105.9617	0.6284	105.5815	0.2673
May	105.5000	106.0960	0.5649	105.9036	0.3826
June	106.3553	106.6087	0.2383	106.4780	0.1154

A. Selecting Samples

In order to compare the forecasting performances of the proposed model with traditional methods, various economic and financial data collected from February 2001 to December 2011 is used as the sample, and the months which data is not complete are excluded. The full samples are divided into a training set (from February 2001 to December 2010), and a testing set (from February 2011 to December 2011).

B. Selecting Variables

The CPI is chosen as the dependent variable. Consider the various factors affecting the consumer price; this paper selects the follow seven major variables: X1, the disposable income of residents; X2, fixed asset investment; X3, products price index; X4, price index of agricultural production; X5, money supply; X6, bank interest rates (personal current interest rates); X7, exchange rate (the RMB against the U.S. dollar) [19]. In order to unify the dimensionless, these variables are unified to "rate" to measure their magnitude. In addition, considering the hysteresis characteristics of the price index, the lag order is set to 2, plus two lagged variables: X8, the first-order lag of CPI; X9, the second-order lag of CPI.

C. Selecting an External Criterion

Minimum of the estimated residual is selected as external criterion.

$$\min[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2]$$

The sample is analyzed by GMDH method and the GMDH modeling based on polynomial spline estimation. The eleven months CPI of the testing set are forecasted and tabulated in Table I, and are shown in Fig. 1.

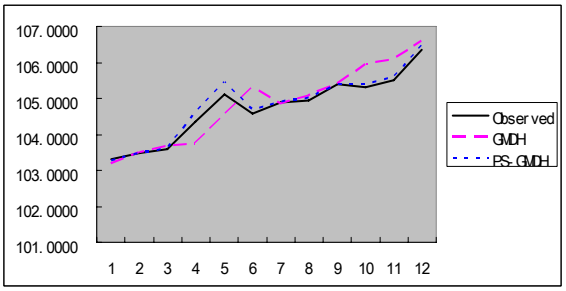

Fig. 1 Comparison of the forecasted results of GMDH, SP-GLSSVM models for CPI.

As shown in Fig. 1, the relative error of SP-GMDH is smaller than the GMDH models. It is evident that the SP-GMDH model performed better than the GMDH models in the testing process.

VI. Conclusions

The GMDH algorithm can fully exploit the real internal structure of the studied object. Layers automatically screening of the model structure and variables in the modeling process can ensure the convergence speed of computation. Non-parametric method does not require pre-specifying the functional relationship between the variables. It greatly reduces the influence of subjective factors. Polynomial spline estimation used the spline function to simulate the variation between the variables. The method predicts the value of the variable, and the convergence speed can reach the global optimum. In this paper the polynomial spline function used to instead the transfer function of GMDH to characterize the relationship between the input variables and output variables. It has proved that the estimators of spline function achieved the global optimal convergence rate. This rate is faster than the convergence rate in the external point of the Kernel estimation. And this rate maintain globally consistent. Therefore polynomial spline estimation has better fitting results and forecasting functions than the non-parametric kernel estimation.The results from the illustrative case show that the new method can forecast CPI in more accurate matter.

References

[1] Ivakhnenko A. G. Heuristic self-organization on problems of engineering cybernetics. *Automatic*. 1970, 6(3), pp. 207-219.
[2] Liu G Z,Yan K Q,Kang Y L. GMDH- type Neural Network Algorithm and its Application. *Mathematics in Practice and Theory*. 2001, 31(4), pp. 464-469.
[3] He C Z, Lv J P. Study of Self-organizing Data Mining Theory and the Complexity of Economic Systems. *Systems Engineering -Theory & Practice*. 2001, 12, pp. 1-5.
[4] Johann Adolf Muller, Frank Lemke. *Self-Organizing Data Mining*. Libri Books. Dresden, Berlin.2000, pp. 67-110.
[5] Godfrey C. Onwubolu. Design of hybrid differential evolution and group method of data handling networks for modeling and prediction. *Information Sci*. 2008,178, pp. 3616– 3634.
[6] Petr Buryana and Godfrey C. Onwubolub. Design of enhanced MIA-GMDH learning networks. *International Journal of Systems Science*. 2001, 42(4), pp. 673-693.
[7] Tian Y X, Tan D J. GMDH Modeling for Forecasting Based on Local Linear Kernel Estimation. *Journal of Systems Engineering*. 2008, 23(1), pp. 9-15.
[8] Meysam Shaverdi, Saeed Fallahi, Vahhab Bashiri. Prediction of Stock Price of Iranian Petrochemical Industry Using GMDH-Type Neural Network and Genetic algorithm. *Applied Mathematical Sciences*, 2012, 6(7), pp. 319 – 332.
[9] Ye A Z. *Nonparametric Econometrics*. Nankai University Press. Tianjin. 2003, pp. 82-92.
[10] Jianhua Z.Huang and Haipeng Shen. Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *Scandinavian Journal of Statistics*, 2004, 31(4), pp. 515-534.
[11] De Boor C. *A practical guide to splines*. Springer. New York. 1978, pp. 168-203.
[12] Brown L D, Levine M. Variance Estimation in Nonparametric Regression via the Difference Sequence Method . *The Annals of Statistics*. 2007, 35(5), pp. 2219—2232.
[13] Cai T T, Levine M, Wang L. Variance Function Estimation in Multivariate Nonparametric Regression with Fixed Design . *Journal of Muhivariate Analysis*. 2009, 100(1), pp. 126—136.
[14] Huang, J. Z, Shen. H. P. Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *Scandinavian Journal of Statistics*.2004, 31(4), pp. 515-534.
[15] Wu X Q, Tian Z, Li X B. Spline Estimates in Functional-Coefficient Linear Autoregressive Models. *Journal of Mathematical Reseach and Exposition*. 2007, 27(4), pp. 869-875.
[16] Stone, C. J., Hansen, M., Kooperberg, C. & Truong, Y. Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*. 1997, 25, pp. 1371– 1470.
[17] Huang, J. Z. Projection estimation in multiple regression with applications to functional ANOVA models. *The Annals of Statistics*. 1998, 26, pp. 242 – 272.
[18] Huang, J. Z. Concave extended linear modeling: a theoretical synthesis. *Statist. Sinica*. 2001, 11, pp. 173– 197.
[19] Liu H B, Liu L. Analysis on the Influencing Factors of CPI Based on VAR Model. *Journal of Yunnan University of Finance and Economics*. 2009, 1, pp. 119-124.