CSCanada
International Business and Management
Vol. 16, No. 1, 2018, pp. 23-28
DOI:10.3968/10092

ISSN 1923-841X [Print]
ISSN 1923-8428 [Online]
www.cscanada.net
www.cscanada.org

A Novel Design of Hybrid Polynomial Spline Estimation and GMDH Networks for Modeling and Prediction

LI Qiumin[a],*

[a]Associate Professor. School of Statistics, Chengdu University of Information Technology, Chengdu, China.
*Corresponding author.

Supported by the open project of statistical information technology and data mining key open laboratory of National Bureau of Statistics of China. (No. SDL201603). This work is supported by the introduction of talent research projects of Chengdu University of Information Technology (No. KYTZ201641).

Received 11 July 2018; accepted 20 September 2018
Published online 26 September 2018

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 novel design of hybrid polynomial spline estimation and GMDH. 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 tax revenue.

Key words: Spline; GMDH; Nonparametric; Bias; Forecast

Li, Q. M. (2018). A Novel Design of Hybrid Polynomial Spline Estimation and GMDH Networks for Modeling and Prediction. *International Business and Management, 16*(1), 23-28. Available from: http://www.cscanada.net/index.php/ibm/article/view/10092
DOI: http://dx.doi.org/10.3968/10092

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 Ivakhnenko in 1970 (Ivakhnenko, 1970, pp.207-219). The GMDH algorithm uses advantages of both self-organizing principle and multilayer neural networks to select best relationships between variables. The main idea of GMDH is the use of feed-forward networks based on short-term polynomial transfer functions combined with emulation of the self-organizing activity behind NN structural learning (Farlow, 1984). J. A. Muller and Frank Lemke (2000) developed and improved self-organizing data mining algorithms. Further enhancements of the GMDH algorithm have been realized in the "KnowledgeMiner" software.

The GMDH algorithm has gradually become an effective tool in many fields such as modeling, forecasting, and decision support and pattern recognition of complex systems, data mining, intelligent classification. The GMDH method has been successfully applied in economy, climate, finance, ecology, medicine, manufacturing and military systems. (Abdel-Aal, Elhadidy, & Shaahid, 2009, pp.1686–1699; Lin, 2012, pp.6665-6671; Dorn, Braga, Llanos, & Coelho, 2012, pp.12268-12279; Mehrara, Moeini, Ahrari, & Erfanifard, 2015, pp.5401-5401).

GMDH provides for a systematic procedure of system modeling and prediction. But in order to improve the forecasting accuracy of GMDH, there are many scholars have done a lot of works. 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. Petr Buryana and Godfrey C. Onwubolu (2001) present an enhanced multilayered iterative algorithm-group method of data handling-type network. 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. 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 is inherently noisy and non-stationary. Zhang Mingzhu, He Changzheng and Liatsi Panos (2012, 2013) brought concept of diversity into GMDH called D-GMDH to improve the noise-immunity ability. The results show that D-GMDH has good prediction accuracy and is an effective means for economic forecasting when data is contaminated by noise. Li Qiumin, Tian Yixiang and Zhang Gaoxun (2014, pp.2301-2308) centers on a new GMDH (Group Method of Data Handling) algorithm based on the k-nearest neighbor (k-NN) method. Instead of the transfer function that has been used in traditional GMDH, the k-NN kernel function is adopted in the proposed GMDH to characterize relationships between the input and output variables.

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

1. THE FUNDAMENTAL OF GROUP METHOD OF DATA HANDLING (GMDH) MODEL

GMDH has been applied to a host of 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.

$$y=f(x_i, x_j) \tag{1}$$

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:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j\geq i}^{m} a_j x_i x_j + \sum_{i=1}^{n}\sum_{j\geq i}^{m}\sum_{k\geq j}^{m} a_{ijk} x_i x_j x_k + \mathrm{L} \tag{2}$$

Where y is the output variable 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.

2. THE POLYNOMIAL SPLINE ESTIMATION

"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. (Carl De Boor, 1978; Huang & Shen, 2004, pp.515-534; Brown & Levine, 2007, pp.2219-2232).

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:

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

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

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

Minimize

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

Have the estimated value $\hat{\beta}_i (i=1,2,\cdots,M+4)$ of β_i,

$\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_{M+4})^T$, $Y=(Y_1,Y_2,\cdots,Y_n)^T$,

$$\hat{\beta}=(W^TW)^{-1}W^TY \tag{6}$$

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

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

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*.

3. 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).

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

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. 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, $\| \hat{m} - m \|_2 = o_p(1), j = 1,2,\cdots,n$.

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

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

Theorem 2. Assume that the conditions A1-A4 hold, and $l = 2$, the estimators of polynomial spline function $m(x)$ $m(x)$ can achieve the global optimal convergence rate $O_p(n^{-2/5})$.

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

The global optimal convergence rate is $O_p(n^{-2/5})$ (Stone, Hansen, Kooperberg, and Truong, 1997,, pp.1371-1470; Huang, 1998, pp.242-272; Huang, 2001, pp.173-197; Huang and Shen, 2004, pp.515-534). 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.

Theorem 3. Assume that the conditions A1-A4 holds, and $l = 3$, 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.

4. AN ILLUSTRATIVE CASE

A tax is a financial charge imposed upon a taxpayer by a government in order to meet public needs. Taxes are an important indicator of the national economy, and a comprehensive reflection of economic situation. Tax revenues are affected by many factors, such as the level of economic development, the design of the tax system, the scope of government functions, etc. Ye lin (2006, pp. 251-255) apply GMDH and BP neural network to tax forecasting, choose the factors as: GDP, the added value of the first, second, third industry, investment in fixed assets, total volume of imports and exports, industrial added value, industrial sales revenue. Chang and Liu (2007, pp.1653-1654) select the added value of the second, third industry, Investment in fixed assets, Total volume of imports and exports to forecast tax based on SVM according to the gray correlation analysis results.

Based on the above findings of scholars, this paper selects the follow six major variables: X1, GDP; X2, the added value of the second industry; X3, the added value of the third industry; X4, investment in fixed assets; X5, Total volume of imports and exports; X6, industrial added value. Output variable Y is tax revenue. The various economic and financial data collected from the first quarter of 2009 to the fourth quarter of 2017 is used as the sample. The full samples are divided into a training set (from 2009 to 2015), and a testing set (form 2016 to 2017). Data is from the China Statistical Yearbook.

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 eight quarters tax of the testing set are forecasted and tabulated in Table 1, and are shown in Figure 1.

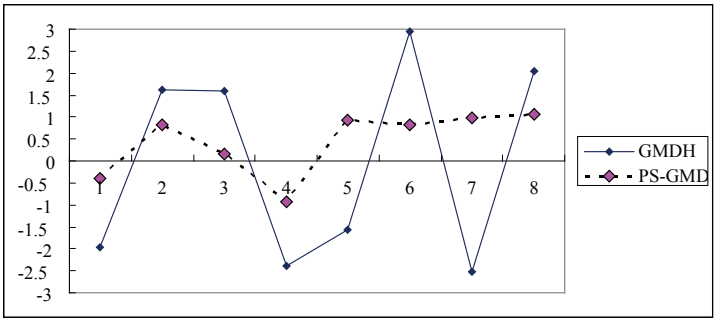

Figure 1
Comparison of the Relative of GMDH, PS-GMDH Models for Tax

Table 1
Tax, Forecast Results and Relative Error (2016.1-2017.4)

Quarter	tax	GMDH forecasting	Relative error (%)	PS-GMDH forecasting	Relative error (%)
2016.1	23438.85	22977.10	-1.97	23342.751	-0.41
2016.2	26589.58	27022.99	1.63	26810.274	0.83
2016.3	21263.75	21602.06	1.59	21299.898	0.17
2016.4	18428.13	17989.54	-2.38	18256.748	-0.93
2017.1	25857.81	25451.84	-1.57	26095.702	0.92
2017.2	29073.82	29928.59	2.94	29315.133	0.83
2017.3	22478.45	21909.75	-2.53	22698.739	0.98
2017.4	23190.80	23666.21	2.05	23438.942	1.07

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

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 tax revenue in more accurate matter.

REFERENCES

Abdel-Aal, R., Elhadidy, M., & Shaahid, S. (2009). Modeling and forecasting the mean hourly wind speed time series using GMDH-based abductive networks. *Renewable Energy, 34*(7), 1686–1699.

Brown, L. D., & Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics, 35*(5), 2219-2232.

Buryana, P., & Onwubolu, G. C. (2001). Design of enhanced MIA-GMDH learning networks. *International Journal of Systems Science, 42*(4), 673-693.

Carl de Boor. (1978). *A practical guide to splines*. New York: Springer.

Chang, Q., & Liu, Q. (2007). Forecasting model of tax based on SVM. *Computer Engineering and Design, 28*(7), 1653-1654.

Dorn, M., Braga, A. L. S., Llanos, C. H., & Coelho, L. S. (2012). A GMDH polynomial neural network-based method to predict approximate three-dimensional structures of polypeptides. *Expert Systems with Applications, 39*(15), (2012), 12268-12279.

Farlow, S. J. (1984). Self-organizing methods in Modeling. New York and Basel: Marcel Dekker.

Huang, J. Z. (1998). Projection estimation in multiple regression with applications to functional ANOVA models. *The Annals of Statistics, 26*, 242-272.

Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statist Sinica, 11*, 173-197.

Huang, J. Z., & Shen, H. P. (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian Journal of Statistics, 31*(4), 515-534.

Ivakhnenko, A. G. (1970). Heuristic self-organization on problems of engineering cybernetics. *Automatic, 6*(3), 207-219.

Li, Q. M., Tian, Y. X., & Zhang, G. X. (2014). The k-nearest neighbour-based GMDH prediction model and its applications. *International Journal of Systems Science, 45*(11), 2301-2308.

Lin, J. S. (2012). A novel design of wafer yield model for semiconductor using a GMDH polynomial and principal component analysis. *Expert Systems with Applications, 39*(8), 6665-6671.

Mehrara, M., Moeini, A., & Ahrari, M. (2015). Investigating the efficiency in oil futures market based on GMDH approach. *Expert Systems with Applications, 42*(12), 5401-5401.

Muller, J. A., & Lemke, F. (2000). *Self-organizing data mining*. Dresden, Berlin: Libri Books.

Onwubolu, G. C. (2008). Design of hybrid differential evolution and group method of data handling networks for modeling and prediction. *Information Science, 178*, 3616-3634.

Shaverdi, M., Fallahi, S., & Bashiri, V. (2012). Prediction of stock price of iranian petrochemical industry using GMDH-Type neural network and genetic algorithm. *Applied Mathematical Sciences, 6*(7), 319-332.

Stone, C. J., Hansen, M., Kooperberg, C., & Truong, Y. (1997). Polynomial splines and their tensor products in extended

linear modeling (with discussion). *The Annals of Statistics, 25*, 1371-1470.

Tian, Y. X., & Tan, D. J. (2008). GMDH modeling for forecasting based on local Linear Kernel estimation. *Journal of Systems Engineering, 23*(1), 9-15.

Ye, L. (2006). Application research on tax forecasting based on GMDH and BP artificial neural network theory. *Mathematics in Practice and Theory, 36*(7), 251-255.

Zhang, M. Z., He, C. Z., & Panos, L. (2012). A D-GMDH model for time series forecasting. *Expert Systems with Applications, 39*(5), 5711-5716.

Zhang, M. Z., He, C. Z., & Panos, L. (2013). D-GMDH: A novel inductive modelling approach in the forecasting of the industrial economy. *Economic Modelling, 30*(1), 514-520.