



Analysis of partial least squares algorithm based on SBM-DEA

Jianqiang Du¹, Zhulin Hao¹, Guolong Wang¹, Riyue Yu², Bin Nie^{1*} and Wangping Xiong¹

¹School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang, China

²School of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang, China

ABSTRACT

The T^2 ellipse assisted analysis method of the partial least squares (PLS) is able to recognize the noise, however, it fails to analyze the noise in the multidimensional space. In the paper, we propose the slacks-based measure (SBM) algorithm to optimize PLS. Firstly, evaluating the sample data comprehensively with SBM, we can gain the valid data. Secondly, analyzing the data based on the PLSR. The two steps can avoid the impact which the noise data have on the regression accuracy and make up the aided analysis technology of the PLSR. Through the calculation of traditional Chinese medicine (TCM) experiments, for two dependent variables, we find out that the average relative errors of the optimized PLS with SBM algorithm are 5.0844% and 8.7485%, which are lower than the results (5.5825% and 9.2810%) by using the PLSR. Besides, for a single dependent variable of the data of tool wear test, the average relative error optimized by SBM is 2.6984%, which is lower than 3.3526% calculated by utilizing the PLS. The experiments result indicate that the regression precision of the PLSR optimized by SBM is much higher than PLSR.

Key words: DEA; slacks-based measure; PLS; auxiliary analysis technology

INTRODUCTION

Multiple regression analysis is a conventional regression statistical analysis method, but it need the data samples meet the Gauss Markov assumptions using the principle of the least squares. However, multiple regression doesn't meet the Gauss Markov assumptions[1], when data samples point is less than the number of variables and independent variables exist serious multicollinearity it doesn't work at this moment. Therefore, Krishnan[2] extends and expands the partial least squares method (PLS) which combines the principal component analysis and canonical correlation analysis with multiple regression analysis; The partial least squares (PLS) is analyzed in the role of eliminating multicollinearity by Guo Jianxiao[3] in detail and is verified that can well eliminate the impact of multicollinearity in the data samples, he elaborates the aided analysis of PLS, points out that the T^2 ellipse aided analysis can be used to determine the importance of data samples points and identities specific points. But the T^2 ellipse aided analysis isn't unable to identify the samples point which is important when the principal component dimension is increased to more than three dimensional space, at this time, all auxiliary analysis technology and method which partial least squares possesses don't identify the wrong data. Accordingly, based on the limitations existing in the analysis of multidimensional space and the problem which wrong data gets error model, Du[4] deepens the SBM model integrally and confirms the feasibility about identifying the specific samples by using the data samples as the decision making unit to calculate the efficiency value; Through the multivariate linear regression method to reduce the dimension of input index in the Data Envelopment Analysis(DEA) of economic theory, Ma et.al[5] make the DEA efficiency be further specified and get good result. However, the traditional DEA model (such as BCC and CCR) can't reflect the characteristics of the data because of not considering the slackness of the input and output. In view of this, the SBM algorithm is employed to auxiliary analysis technology of the partial least squares to optimize further the partial least squares regression modeling method in this paper.

THE SBM-DEA ALGORITHM

Data envelopment analysis put forward by Charnes A[6] can be used to estimate the relative efficiency of the decision making unit with multiple inputs and multiple outputs. The traditional DEA (such as CCR and BCC) ensures the boundary of the relative effectiveness or the convex of the indifferent curve, but it may lead to the congestion and slacks of the inputs[7]. It is more difficult for all the each decision making unit the overall data samples to evaluate the relative effectiveness, when increasing the input or output and considering the corresponding slack. Kaoru proposes SBM[8] of efficiency in Data Envelopment Analysis. This scalar measure deals directly with the output shortfalls and the input excesses of the decision making unit (DMU) concerned, so it is a effective method to solve the slacks issue[9].

We will dispose n DMUs with s_2 undesirable outputs, s_1 good outputs and m input matrices, $Y^b = (y_{ij}) \in R^{m \times s_2}$, $Y^g = (y_{ij}) \in R^{m \times s_1}$ and $X = (x_{ij}) \in R^{m \times n}$. Suppose that the data samples are positive, i.e. $X > 0$ and $Y > 0$. Hence, SBM model is shown by the following:

$$\rho = \min \frac{1 - \frac{1}{m} \sum_{i=1}^m \frac{s_i^-}{x_{i0}}}{1 + \frac{1}{s_1 + s_2} \left(\sum_{r=1}^{s_1} \frac{s_r^g}{y_{r0}^g} + \sum_{r=1}^{s_2} \frac{s_r^b}{y_{r0}^b} \right)} \quad (1)$$

$$s.t. \begin{cases} \lambda X + s^- = x_0 \\ \lambda Y^g - s^g = y_0^g \\ \lambda Y^b + s^b = y_0^b \\ s^- \geq 0, s^g \geq 0, s^b \geq 0, \lambda \geq 0 \end{cases}$$

where ρ is an effort to estimate the efficiency, the vectors $s^- \in R^m$, $s^g \in R^{s_1}$ and $s^b \in R^{s_2}$ indicate the input excess and output shortfall in this expression, respectively, and are called slacks. For the particular decision making unit evaluated as follows:

Definition 1:

- (1) A DMU is SBM-efficient if $\rho=1$ which is equipment to $s^g = 0$, $s^b = 0$ and $s^- = 0$;
- (2) A DMU is weak effective but close to the effective, existing the necessity of the input and output improvement. The problem given above is a nonlinear programming because of containing the nonlinear term. To facilitate the use of MATLAB programming calculation, utilizing the Cooper transformation[10], we can transform it into a linear program as follows:

$$\rho^* = \min t - \frac{1}{m} \sum_{i=1}^m \frac{S_i^-}{x_{i0}} \quad (2)$$

$$s.t. \begin{cases} t + \frac{1}{s_1 + s_2} \left(\sum_{r=1}^{s_1} \frac{S_r^g}{y_{r0}^g} + \sum_{r=1}^{s_2} \frac{S_r^b}{y_{r0}^b} \right) = 1 \\ \lambda^* X + S^- = x_0 t \\ \lambda^* Y^g - S^g = y_0^g t \\ \lambda^* Y^b + S^b = y_0^b t \\ S^- \geq 0, S^g \geq 0, S^b \geq 0, \lambda^* \geq 0, t \geq 0 \end{cases}$$

$$\rho = \rho^*, \lambda = \lambda^*/t, s^- = S^-/t, s^g = S^g/t, s^b = S^b/t$$

A NOVEL APPROACH BASED ON SBM-DEA

SBM model can be directly used to evaluate the relative efficiency of the original data samples, not needing any data preprocessing[11]. We screen the valid samples according to the effective value of the evaluation, then establishing the model about partial least squares regression (PLSR) method. The concrete steps are as follows:

(1) Suppose that $S = (s_1, s_2, \dots, s_m)$ is the data samples, m is the number of samples, independent variables and dependent variables are (x_1, x_2, \dots, x_p) and (y_1, y_2, \dots, y_q) . According to the SBM model, (s_1, s_2, \dots, s_m) is took as decision making unit, called stations of the data samples; (y_1, y_2, \dots, y_r) is the undesirable outputs not being expected to increase; $(y_{r+1}, y_{r+2}, \dots, y_q)$ is the good output being expected to increase; (x_1, x_2, \dots, x_p) is the input. Then, (2) is adopted to solve ρ of the decision making unit. For a sample, we call the sample as the effective sample if $\rho = 1$, namely, the effective value of the SBM model take the threshold for 1 as a standard. Therefore, the new selected data samples are analyzed by utilizing partial least squares method. The concrete SBM algorithm workflow is as follows:

Input : $S = (s_1, s_2, \dots, s_m)$

Output : ρ

Initialize :

Repeat

Compute ρ Using the formulation (2)

Until converge

Result :

$$\rho = \rho^*, \lambda = \lambda^*/t, s^- = S^-/t, s^g = S^g/t, s^b = S^b/t$$

(2) The effective data sample selected through step(1) is preprocessed. Then we get the processing data matrix: The independent variables are $X = (x_1, \dots, x_i, \dots, x_p)$; The dependent variables are $Y = (y_1, \dots, y_j, \dots, y_q)$, where n is the number of the data samples, p is the number of the independent variables, q is the number of the dependent variables.

(3) Let t_1 and u_1 be the first principal component of X and Y , i.e. $t_1 = Xw_1, u_1 = Yv_1$, where w_1 and v_1 are the first axis of X and Y , i.e. $\|v_1\|=1, \|w_1\|=1$, these axes are the unit column vector. t_1 and u_1 must meet the following two conditions[12]:

1) the variation information is the largest: $Var(t_1) \rightarrow \max, Var(u_1) \rightarrow \max$

2) the degree of correlation is the biggest too: $r(t_1, u_1) \rightarrow \max$

It can get that the covariance is the maximum synthetically: $Cov(t_1, u_1) = r(t_1, u_1)\sqrt{Var(t_1)Var(u_1)} \rightarrow \max$

Then basing on the Lagrange Algorithm, we can obtain: $X = t_1 p_1^T + X_1$ $Y = t_1 r_1^T + Y_1$

where X_1 and Y_1 are the residuals information matrix of X and Y , the regression coefficient vector p_1 and r_1 are as follows:

$$\begin{cases} p_1 = \frac{X^T t_1}{\|t_1\|^2} \\ r_1 = \frac{Y^T t_1}{\|t_1\|^2} \end{cases} \quad (3)$$

(4) Replacing X and Y with the residuals information matrix, X_1 and Y_1 , then count t_2 and t_2 of the second component as well as w_2 and v_2 of the second axle, namely, $t_2 = X_1 w_2, u_2 = Y_1 v_2$. Then it can be inferred that $X_1 = t_2 p_2^T + X_2$ and $Y_1 = t_2 r_2^T + Y_2$, and the regression coefficient vector p_2 and r_2 are as follows:

$$\begin{cases} p_2 = \frac{X_1^T t_2}{\|t_2\|^2} \\ r_2 = \frac{Y_1^T t_2}{\|t_2\|^2} \end{cases} \quad (4)$$

(5) So we recycle residuals information matrix to calculate iteratively, and assume the rank of X to be m (if it has A principal component, $A \leq r(X) = m$), then it has:

$$\begin{cases} X = t_1 p_1^T + t_2 p_2^T + \dots + t_m p_m^T + X_m \\ Y = t_1 r_1^T + t_2 r_2^T + \dots + t_m r_m^T + Y_m \end{cases} \quad (5)$$

t_1, t_2, \dots, t_m are the linear combination of $\{x_1, x_2, \dots, x_p\}$, and among them, X_m, Y_m is the residuals information matrix of the number m .

(6) Reducing the equation, and according to the properties of the PLS, it has:

$w_h^* = \prod_{k=1}^{h-1} (E - w_k p_k^T) w_h$ & $t_h = X w_h^*$, then we obtain:

$$Y = X \left(\sum_{i=1}^m w_i^* r_i^T \right) + Y_m \quad (6)$$

In the regression equation (6), We order the regression coefficient vector $B = \sum_{i=1}^m w_i^* r_i^T$, then it has

$$Y = XB + F_m \quad (7)$$

(7) In the process of PLS, because the follow-up principal components fail to provide more significant information to explain Y , utilizing more will destruct the statistical trend of the regression model and lead to the wrong conclusion. For PLS, it is not necessary for the whole principal components to structure regression model. According to the size of sample data sets, for small sample data sets, we adopt the leave one outcross validation to judge the number of the valid principal components [13], and the calculation formula of cross validity Q_h^2 is as follows:

$$Q_h^2 = 1 - \frac{\sum_{j=1}^q PRESS_{hj}}{\sum_{j=1}^q SS_{(h-1)j}} = 1 - \frac{PRESS_h}{SS_{h-1}} \quad (8)$$

The criterion of the leave one outcross validation:

1) If $Q_h^2 \geq 1 - 0.95^2 = 0.0975$, it is valid to add the t_h component, and the model can be improved dramatically.

2) If $\exists k \in \{1, 2, \dots, q\}$, it has $Q_h^2 \geq 0.0975$.

The processes of the concrete algorithm are as follows:

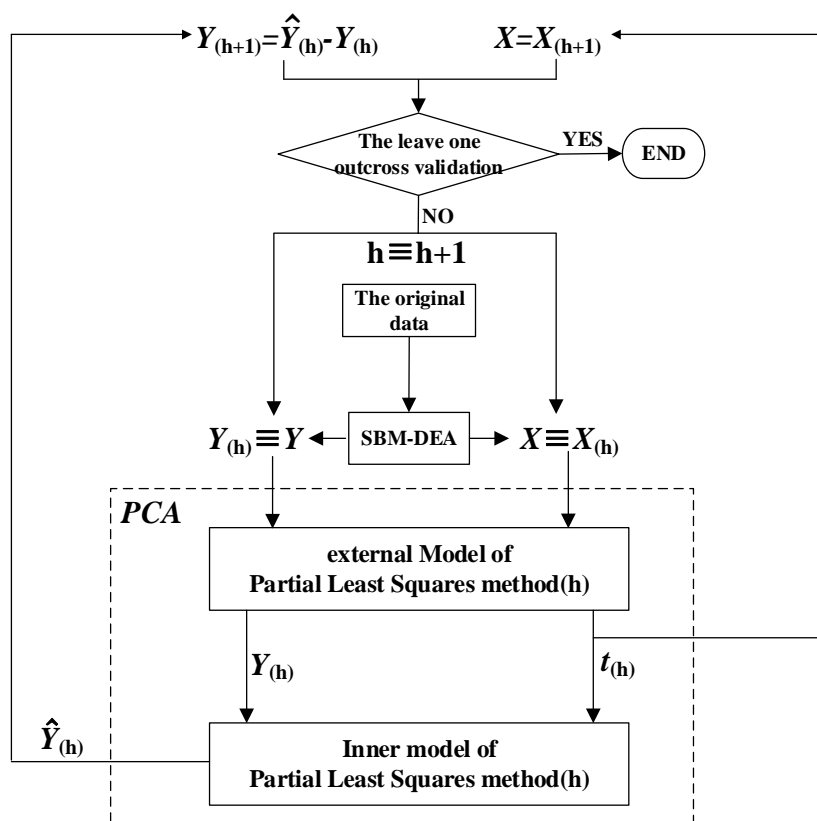


Fig.1:A novel approach based on SBM-DEA

RESULTS

Rough In order to compare the effect of the partial least squares regression optimized by SBM method with the simple partial least squares method, we analyze the experimental data of traditional Chinese Medicine and the tool wear respectively. This article refers to the data of the effect of the dachengqi decoction and its components on intestinal blood flow and the perimeter of obstruction in the rat, which comes from Key Laboratory of Modern Preparation of TCM, Ministry of Education of Jiangxi University of traditional Chinese Medicine. In the table 1, the left column is the experimental prescription species, and along with the primary prescription, the mixture of the experimental prescriptions adopt the formula designed uniformly as well as the consumption of them is discounted by the primary clinical dosage. The first species is the original prescription, and other species are adjusted on the basis of the original prescription. $x_1 \sim x_9$ are the components of rhubarb and $x_{10} \sim x_{12}$ are the contents of cortex magnolia officinalis, while y_1 is the small intestine perimeter from the ligation of 1cm of the obstruction rat and y_2 is the blood flow of the terminal ileum vascular in rats. Therefore, we totally have 12 chemical independent variables and 2 dependent variables. According to the scheme of the experimental data of the traditional Chinese medicine, it shows that a linear relationship between y and x and 2 dependent variables are included in the expected output, then we calculate the efficiency value of each sample observation by MATLAB programming (table 1). Thereout we eliminate the third and seven prescription, because the efficiency values of their sample observation fall below 1. Then we model the partial least squares regression to analyze the remaining 8 samples, because of including small sample data, therefore we extract 2 principal components by adopting leave-one-outcross-validation and get the partial least squares regression equation optimized by SBM algorithm (formula 9 and 10). Because the 3 and 7 prescription are judged as "noise" through utilizing SBM algorithm, however, the experiment is implemented in the same condition, so they should be contained to determine the reliability of PLS model optimized by SBM and calculate the forecast value. After that, we analyze the 10 sample points which are not optimized by SBM in the table 1 through the PLS directly. As above, according to leave-one-outcross-validation, 2 principal components should be distilled. In order to compare with the above experiment, we find out the relative forecasts of 2 dependent variables and the relative errors of them as well as the relative average errors of all the sample forecasts. Finally, we get the table 2.

Table 1 The effect of the dachengqi decoction and its components on intestinal blood flow

NO.	Total anthraquinone			Combined anthraquinone					Magnolol acid			Magnolol		y_1	y_2	value
	Aloe Emodin	Rhein	Chryso-phenol	Emodin	Rhein	Chryso-phenol	Emodin	Rhein	Honokiol	Emodin	ether					
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}				
1	0.0625	0.0468	0.0945	0.0724	0.0265	0.0455	0.0324	0.0213	0.0626	0.0226	0.0138	0.0072	2.613571	1		
2	0.045	0.0317	0.0558	0.0899	0.0214	0.0122	0.0107	0.0066	0.0649	0.0138	0.0134	0.0164	2.242653	1		
3	0.0075	0.0085	0.0126	0.0139	0.0063	0.0062	0.007	0.0062	0.0111	0.0047	0.016	0.0213	2.114380	1		
4	0.035	0.0278	0.0434	0.0532	0.0155	0.0245	0.0204	0.0164	0.048	0.014	0.0161	0.0239	2.444747	0.50		
5	0.018	0.0097	0.0232	0.0159	0.0036	0.0128	0.0071	0.0139	0.0135	0.0031	0.0122	0.0179	2.453783	1		
6	0.034	0.0233	0.0631	0.0654	0.0184	0.0215	0.0155	0.025	0.0586	0.0162	0.0085	0.0117	2.394155	1		
7	0.0227	0.0104	0.03195	0.0213	0.0478	0.0047	0.007	0.0154	0.018	0.0037	0.003	0.0032	2.592664	1		
8	0.1006	0.0875	0.1841	0.2119	0.068	0.0509	0.071	0.0933	0.1973	0.0625	0.014	0.0136	2.593956	0.23		
9	0.106	0.096	0.1982	0.1701	0.0495	0.0717	0.0701	0.0695	0.1504	0.042	0.0079	0.0045	2.313472	1		
10	0.054	0.0441	0.0871	0.0998	0.0277	0.0383	0.0313	0.023	0.0918	0.0243	0.0042	0.0133	2.493244	1		

$$y_1 = 0.1872x_1 - 0.2312x_2 - 0.0919x_3 - 0.1022x_4 + 2.5421x_5 - 0.2728x_6 - 0.5223x_7 - 0.8104x_8 - 0.1418x_9 - 0.7556x_{10} - 10.5826x_{11} - 7.8731x_{12} + 2.6045 \quad (9)$$

$$y_2 = -761.9823x_1 + 608.4151x_2 + 222.8416x_3 + 253.4422x_4 - 8682.2195x_5 + 727.3645x_6 + 1514.6476x_7 + 2435.9388x_8 + 370.4591x_9 + 2158.6555x_{10} + 35620.7456x_{11} + 26628.3447x_{12} + 2852.8319 \quad (10)$$

Table 2 the effect comparison between PLSR optimized by SBM and the simple PLSR in the experimental data of the traditional Chinese medicine

NO.	PLS optimized by SBM				PLS directly				
	The predicted value		The relative error		The predicted value		The relative error		
	y_1	y_2	y_1	y_2	y_1	y_2	y_1	y_2	
1	2.3814	3532.3327	0.0676	0.0166	2.4336	3630.1301	0.0876	0.0108	
2	2.3409	3696.2619	0.0638	0.4435	2.3829	3829.6590	0.0450	0.3932	
3	2.2649	3985.5097	0.0957	0.0490	2.3120	4165.3636	0.0734	0.0901	
4	2.2281	4078.2398	0.0497	0.0910	2.3187	4314.9546	0.0869	0.1409	
5	2.3183	3799.7989	0.0399	0.0284	2.3521	3890.3811	0.0538	0.0044	
6	2.4030	3483.0051	0.0165	0.1602	2.4294	3489.3104	0.0054	0.1617	
7	2.6431	2697.6333	0.0239	0.0245	2.5282	2598.8030	0.0205	0.0126	
8	2.2804	3774.4309	0.0474	0.0007	2.4671	3953.0409	0.1195	0.0459	
9	2.4078	3363.2161	0.0947	0.0239	2.5287	3389.1116	0.0423	0.0313	
10	2.4307	3364.2423	0.0092	0.0371	2.4671	3364.2457	0.0238	0.0371	
The relative average errors			5.0844%	8.7485%				5.5825%	9.2810%

Table 3 The sample data of the tool wear test

NO.	x_1	x_2	x_3	x_4	x_5	y	SBM
1	482.86	751	620.66	3.438	102.4	58	1
2	494.49	839	665.88	3.655	307.2	134	1
3	545.95	957	701.96	4.293	512	177	1
4	499.07	923	708.18	4.052	614.4	185	0.9828
5	398.54	745	595.75	3.544	716.8	186	1
6	443.4	781	691.43	3.721	819.2	188	0.8664
7	475.45	874	685.48	3.935	921.6	208	0.8842
8	478.01	927	761.19	4.026	1126.4	254	1
9	517.2	1069	800.1	4.46	1228.8	276	1
10	513.44	1064	822.9	4.326	1331.2	290	1

Then we utilize the same scheme to analyze the data of tool wear[14] (table 3), and reject the 4, 6 and 7 sample with the efficiency values falling below 1. According to the tool wear test, y and x_1 , x_2 , x_3 , x_4 , x_5 have a nonlinear relationship, and referring to the eliminated sample data, we execute the logarithmic transformation and the partial least squares regression. On the basis of the leave-one-out cross-validation, we confirm 5 principal components should be picked up. After getting the equation, we carry out antilog transformation and then attain the relation equation between y and x (formulation 11). Afterwards, we analyze the 10 sample points in the table 3 by utilizing the PLS directly. As above, 3 principal components are confirmed, and we calculate the relative forecasts of variables, the relative errors of them as well as the relative average errors, then compare them.

$$y = 0.0002983x_1^{2.2427}x_2^{1.3656}x_3^{-1.7361}x_4^{-2.7593}x_5^{0.8319} \quad (11)$$

Table 4 the effect comparison between PLSR optimized by SBM and the simple PLSR in the data of the tool wear test

NO.	PLS optimized by SBM		PLS directly	
	The predicted value	The relative error	The predicted value	The relative error
1	58.1546	0.0027	60.6156	0.0451
2	133.1242	0.0065	126.2826	0.0576
3	177.8933	0.0050	171.2658	0.0324
4	186.3175	0.0071	183.4204	0.0085
5	186.5204	0.0028	181.6716	0.0233
6	190.4888	0.0132	184.9561	0.0162
7	249.2811	0.1985	230.6211	0.1088
8	252.9336	0.0042	247.6665	0.0249
9	272.2765	0.0135	277.4667	0.0053
10	294.7194	0.0163	293.8189	0.0132
The relative average errors		2.6984%	3.3526%	

In the table 2 and 4, using the relative average error as reliability criterion and referring to 2 dependent variables, we figure out the relative average errors by utilizing the PLS optimized by SBM algorithm are 5.0844% and 8.7485%, which is below to the results(5.5825% and 9.2810%)by using the PLS directly. According to the 1 dependent variable of the data of tool wear test, the relative average error optimized by SBM is 2.6984%, which is lower than 3.3526% calculated by the PLS directly.

CONCLUSION

Through the above analysis, we can obtain the following conclusions: Firstly, we propose to utilize SBM to optimize the PLS model, which improves the model precision and calculate the efficiency value of the sample points depending on SBM as well as analyze its characteristics, therefore, the “bad data” can be eliminated better and we obtain more reliable model finally. Secondly, comparing to the alone PLS, the relatively average error of PLSR optimized by SBM is lower. Thirdly, different DEA models have a variable effect on data samples, so we can adopt different DEA models to reduce the influence which invalid data have on the regression model. Fourthly, it is able to pull different DEA models into the regression model of the data of Traditional Chinese Medicine to provide better technical support to the Traditional Chinese Medicine.

Acknowledgments

This work is supported by the Key Laboratory of modern preparation of Traditional Chinese Medicine (TCM), Ministry of education and a national natural science foundation (61363042) and the National Key Basic Research Program of China (973 Program) (2010CB530602), This research also is supported by Jiangxi Province Innovation Foundation for Graduate (YC2013-S226).

REFERENCES

- [1] Xu Qun. The research on non-linear regression analysis methods[D]. Hefei University of Technology, 2009.
- [2] Krishnan A, Williams L J, McIntosh A R, et al. *NeuroImage*. 2011, 56(2): 455-475.
- [3] Guo Jianxiao, Study on Improved High-Dimension and Nonlinear Partial Least-Squares Regression Method and Applications[D]. Tianjin University, 2010.
- [4] Du J, Liang L, Zhu J. *European Journal of Operational Research*. 2010, 204(3): 694-697.
- [5] Ma Shengjun, Wang Dongmei, Ma Zhanxin, et al. *Journal of Inner Mongolia Agricultural University(Natural Science Edition)*. 2012, 33(1): 231-235.
- [6] Cooper W W. Handbook on Data Envelopment Analysis[M]. Springer US, 2011.
- [7] Wei Quanling. Data envelopment analysis model to evaluate the relative effectiveness —— DEA and network DEA[M]. Beijing: China Renmin University Press, 2012.
- [8] Chen J, Deng M, Gingras S. *Computers & operations research*. 2011, 38(2): 496-504.
- [9] Zhou Y, Xing X, Fang K, et al. *Energy Policy*. 2012, 57(0): 68-75.
- [10] Li H, Fang K, Yang W, et al. *Mathematical and Computer Modelling*. 2013, 58(5-6): 1018-1031.
- [11] Tone K. *European Journal of Operational Research*. 2010, 200(3): 901-907.
- [12] Sun Fenglin, Hao Zhifeng. *Computer Engineering and Design*, 2010, 31(12): 2826-2829.
- [13] Rodriguez J D, Perez A, Lozano J A. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2010, 32(3): 569-575.
- [14] Zhang Xiaohai, Jin Jiashan, Gen Junbao. *Journal of Zhejiang University(Engineering Science)*. 2011, 45(9): 1688-1692.