ELSEVIER

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 5025 – 5029

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineering and Information Science

Coal Price Index Forecast by a New Partial Least-Squares Regression

Bo Zhang[1],Junhai Ma[1,2] *

1 College of Management Economic，Tianjin University，Tianjin 300072，China
2 Tianjin University of Finance & Economics, Tianjin 300222, China

Abstract

Deviation of coal price has great influence on growth of China's economic. Daily coal price indexes in Qinhuangdao were collected. Past twenty days were used to predict next day index. The principal components of twenty days were extracted. The function between output variable and components was fitted by linear, quadratic and exponential model. This improved traditional partial least-squares regression. Traditional method such as multivariate linear regression and polynomial regression were coming into comparing with our method. Improved quadratic partial least-squares obtained the smallest relative errors in mean and variance for ten reserved indexes. Those ten errors had minimum 0.3%, median 3.3% and maximum 9.7%. The ideal forecast precision certified that quadratic partial least-squares was suitable for coal price indexes.

Keywords: Coal price index, principal components, partial least-squares regression

1. Introduction

Coal, as principle source of fuel dynamic of industry and source of chemical material and civilian energy as well as a key export commodity, plays an important role in China's economic development since China is one of few countries which take coal as primary energy resource. At present, being great strategetic status in national economic, coal contributes to 70% and 66% of production and consumption of one-time energy and is the main support energy in the long future.

We can see that the price of coal is closely related with the timeSince Data is the only base of prediction, coal price indexes from 2 September 2009 to 2 March 2011were collected from Qinhuangdao Coal Information Network that came from real trade contracts and presents the true market. Based on Phase Space Reconstruction, indexes were forecasted by improved partial least-squares.

* Corresponding author. Tel.:+8615222025960
E-mail address: 1tdgy100512@yahoo.cn,2mjhtju@yahoo.com.cn

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2011.08.934

2. Improved Partial Least- Squares

Given two variables sets of independent variables and dependent ones, the first set i.e. independent variables can be expressed as p-dimensional table $X=(x_1,x_2,...,x_p)$ as well as the second set i.e. dependent ones as q-dimensional table $Y=(y_1,y_2,...,y_q)$ with sample size n. The raw data is $n\times(p+q)$ sample matrix $Z=(X_{n\times p},Y_{n\times q})_{n\times(p+q)}$.

Nonlinear partial least-squares regression researched the nonlinear relation between input and output variables. In order to avoid influence of different dimension, standardization was necessary and let standard matrix $E_0=(E_{01},E_{02},...,E_{0p})_{n\times p}$ and $F_0=(F_{01},F_{02},...,F_{0q})_{n\times q}$ represent dimensionless input variables and output variables respectively. Our modified partial least-squares regression was described as follows.

2.1 Extracting principal components

Firstly estimate the first principal component t_1 of E_0 with $t_1=E_0w_1$, where w_1, unit vector with $\|w_1\|=1$, was the first axis of E_0. In the same way, the first principal component u_1 of F_0 with $u_1=F_0c_1$, where c_1, unit vector with $\|c_1\|=1$, was the first axis of F_0. According to principal components analysis, the first component has the largest variation as well as t_1 and u_1 satisfied following condition

$$Var(t_1)\rightarrow\max$$
$$Var(u_1)\rightarrow\max.$$

Meanwhile, the sets of independent and dependent variables were required to have the strongest correlation as well as t_1 and u_1 had the maximum dependence $r(t_1,u_1)\rightarrow\max$. Covariance could be as a measure of dependence, so it equaled $Cov(t_1,u_1)=\sqrt{Var(t_1)Var(u_1)}\cdot r(t_1,u_1)\rightarrow\max$ Then the problem could be solved by following optimization.

$$\max w_1^TE_0^TF_0c_1,\quad s.t\begin{cases}w_1^Tw_1=1\\ c_1^Tc_1=1\end{cases} \tag{1}$$

Satisfying constrained condition $\|w_1\|^2=1$ and $\|c_1\|^2=1$, the optimized values w_1 and c_1 maximized $w_1^TE_0^TF_0c_1$ .It was solved that w_1 and c_1 were eigenvectors of the largest eigenvalue of $E_0^TF_0F_0^TE_0$ and $F_0^TE_0E_0^TF_0$ respectively.

After getting axis w_1 and c_1, the first principal components was

$$t_1=E_0w_1 \tag{2}$$

$$u_1=F_0c_1 \tag{3}$$

2.2 Nonlinear regression between dependent variables and principal components

The next step was to construct regression equation on principal components as

$$E_0 = t_1 p_1^T + E_1, F_0 = u_1 q_1^T + F_1^*, F_0 = t_1 r_1^T + F_1 \qquad (4)$$

In these three equations, the one between F_0 and u_1 had no use in continuing calculation. The equations were all linear in conventional partial least-squares and its descendable methods. We can modify the Eqn. (4) as nonlinear form

$$E_0 = f_1(t_1) + E_1 \qquad (5)$$

$$F_0 = g_1(t_1) + F_1 \qquad (6)$$

2.3 Nonlinear regression by iteration

The second principal components t_2 and u_2 were computed in the same way by substituting E_0 and F_0 with residuals E_1 and F_1 in Eqn. (5) and Eqn. (6). Nonlinear equation could be calculated continuingly

$$E_1 = f_2(t_2) + E_2 \qquad (7)$$

$$F_1 = g_2(t_2) + F_2 \qquad (8)$$

After cycle computation, we obtained

$$E_0 = f_1(t_1) + f_2(t_2) + \cdots + f_A(t_A) \qquad (9)$$

$$F_0 = g_1(t_1) + g_2(t_2) + \cdots + g_A(t_A) + F_A \qquad (10)$$

Because principal components $t_1, t_2, \cdots, t_A$ were linear combination of $E_{01}, E_{02}, \cdots, E_{0p}$, Eqn. (10) could be restored to

$$y_k^* = g_1[\phi_{t_1}(x_1^*, x_2^*, \cdots, x_p^*)] + g_2[\phi_{t_2}(x_1^*, x_2^*, \cdots, x_p^*)] + \cdots + g_A[\phi_{t_A}(x_1^*, x_2^*, \cdots, x_p^*)] + F_{Ak}, \quad k = 1, 2, \cdots, q \qquad (11)$$

where $x_j^* = E_{0j}$, which could be simplified depending on conditions.

3. Forecasting Results

Our coal price indexes with minimum 1113, maximum 2115 and median 1387, were listed in Fig. 1 from which periodicity was obvious. The underlying system of univariate time series of price indexes was multivariate since many factors had impact on it. The key point was to reconstruct the complicated model from finite univariate data, which also called Phase Space Reconstruction. Its foundation is Taken's Theorem which exemplified that information about higher-dimensional space can be hidden in its one component and it is possible to restore the whole system by one variable.

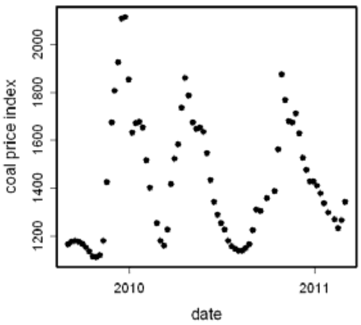


Fig. 1. Coal price indexes in port of Qinhuangdao

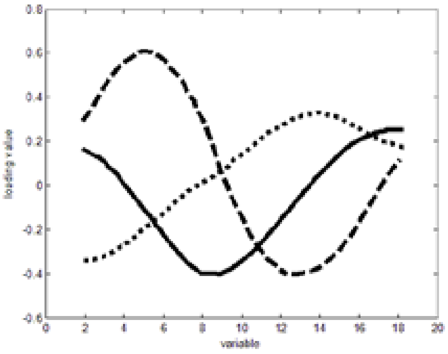
Phase Space Reconstruction was equal to find the function between present value x_t and past ones $x_{t-\tau}, x_{t-2\tau}, \cdots, x_{t-d\tau}$, taking x_t as output variable and $x_{t-\tau}, x_{t-2\tau}, \cdots, x_{t-d\tau}$ as input one. Embedding dimension d and time delay τ could be estimated by Mutual Average Information before fitting.

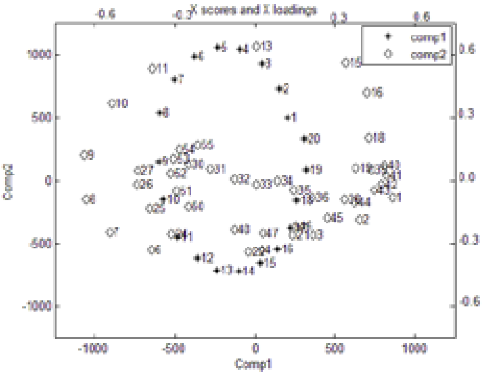
A modified partial least-square, suitable for multivariate dependence variable, set time delay $\tau = 1$ and embedding dimension $d = 20$ large enough.

Mainly computation works were finished by library “pls” in R language.

Ten day indexes, randomly reserved from origin sixty five records, could be for testing effect and the rest for training. No standard transformation was done since all variable came from the same series.

Their scores and loadings illustrated in Fig. 2 and Fig. 3, the first three principal components, most relevant to dependent variable, were calculated according to conventional partial least-squares by traditional orthogonal decomposition. They contributed 36%, 24% and 23% of total variance respectively.





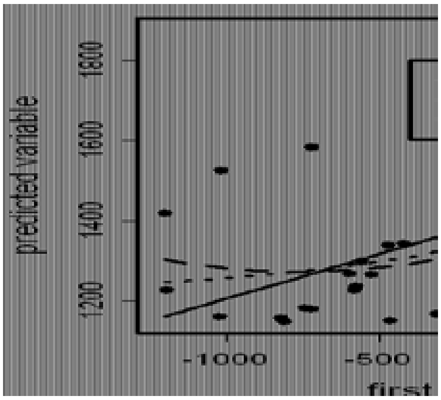


Fig. 3. Scores and loadings of the first and second principle components

Fig. 2 depicted several regression models between output variable and the first principal components, where scatter points was real price indexes, solid line as linear regression model, long dashed curve as quadratic model and shorter dashed curve as exponent model. It was obvious from Fig. 2 that nonlinear regressions such as quadratic and exponent model were superior to linear model.

The fitted parameters of those models were listed as Tab. 1, where quadratic model had the smallest standard error of the three models using the same raw data matrix. Considering standard errors $\hat{\sigma}$ of residuals and R-squared R^2 , quadratic regression was the best, exponent model the better and linear one the worst.

Fitting effects were compared among some kinds of regressions including our modified nonlinear partial least-squares based on the identical sample, and ten reserved data’s absolute relative errors, true value divided by the difference between predicted value and true one, were boxploted In Fig. 3.

Overestimate or underestimate were all thought as errors, so absolute values were compared. In Fig. 3, “line” presented linear partial least-squares, “quad” as quadratic partial least-squares, “exp” as exponent partial least-squares, “ml” as multivariate regression, “pcr” as principal regression, “step” as step polynomial regression and “local” as local polynomial regression.

TABLE I
FITTING EFFECT OF THREE MODELS BETWEEN DEPENDENT VARIABLE AND THE FIRST PRINCIPLE COMPONENTS

Number	Model	Standard Errors $\hat{\sigma}$	R-squared R^2	
			R^2	Adjust R^2
1	Quadratic	131.4	0.6134	0.5949
2	Linear	142.4	0.5347	0.5239

It was seen from Fig. 5 that relative errors of quadratic partial least-squares whose minimum was 0.3%, median 3.3% and maximum 9.7%, were the closest to zero with the lowest height of box which indicated the smallest deviation, which reached ideal forecast. In conclusion, quadratic partial least-squares was appreciated for our coal data.

4. Conclusion

Ignoring nonlinear dependence, conventional linear and nonlinear partial least-squares only considered the linear relation between output variables and principle components, so the model could be chose under condition as well as each principle components were still linear combination of independent variables in our modified partial least-squares. Comparing of regression model, relative errors of reserved data demonstrated quadratic partial least-squares' superiority for coal price index prediction.

5. Acknowledgment

The research work of thesis is fully supported by the National Natural Science Foundation of China (Grant NO. 10772132) and the Doctoral Foundation of Ministry of Education of China (Grant No. 20070056063).

References

[1] Z. Y. Chun, "The Application of Data Mining in the Coal Price Forecast." Master dissertation, Computer Applications Technology, University of Anhui, Anhui, China, 2006.

[2] L. Ming, L. S. Hua, "A coal price forecast model and its application," Journal of Wuhan University of Science and Technology (Natural Science Edition), vol. 30, no. 4, pp. 434–437, Aug. 2007.

[3] H. J. Long, N. Y. Cai, "Prediction of coal price based on Box-Jenkins," China Price, vol. 22, no. 1, pp. 10–11, Jan. 2007.

[4] A. L. Boulesteix, K. Strimmer. "Partial Least Squares: a versatile tool for the analysis of high-dimensional genomic data," Briefings in Bioinformatics, vol. 8, no. 1, pp. 32–44, Jan. 2009.

[5] S. de Jong. "SIMPLS: an alternative approach to partial least squares regression." Chemometrics and Intelligent Laboratory Systems, vol. 18, no. 3, pp. 251–263, Mar. 1993.

[6] M. Momma. "Efficient computations via scalable sparse kernel partial least squares and boosted latent features," In Proceedings of SIGKDD International Conference on Knowledge and Data Mining, SIGKDD´05, Chicago, IL, pp. 654–659.

[7] R. Roman, N. Kramer, "Overview and Recent Advances in Partial Least Squares, "in Subspace, Latent Structure and Feature Selection Techniques, C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor, Ed. German Berlin: Springer, 2006, ch. 2, pp. 34-51.

[8] P. H. Garthwaite, "An interpretation of partial least squares", Journal of the American Statistical Association , vol. 89, no. 425, pp. 122-127, Mar. 1994.

[9] S. de Jong, and C. J. F. ter Braak, "Comments on the PLS kernel algorithm," Journal of Chemometrics, vol. 8, no. 2, pp. 169–174, Mar/Apr. 1994.

[10] B. S. Dayal, and J. F. MacGregor, "Improved PLS algorithms," Journal of Chemometrics, vol. 11, no. 1, pp. 73–85, Jan. 1997.