The application of sparse partial least squares regression in electricity consumption of Yunnan province

Peng Shuting
Kunming University of science and technology, Faculty of science
Kunming, R.P. China
E-mail address:342726185@qq.com

Dai Lin*, Guo Tingting
Kunming University of science and technology, Faculty of science
Kunming, R.P. China
E-mail address:453091926@qq.com

Abstract—**It's extremely important to screen key variables from high-dimesional electricity data that contains many predictors and presents multi-collinearity. In this paper, sparse partial least-squares regression(SPLS) is employed to investigate the electricity consumption from Yunnan province of China. SPLS can automatically select important variables and simultaneously eliminate the uninformative variables. The root mean square errors(RMSE) is used to evaluate the prediction performance and the results show that SPLS is competitive with ordinary least squares (OLS) and partial least squares regression (PLS). In addition, several predictors such as GDP of Yunnan are chosen as key factors with SPLS algorithm.**

Keywords—***partial least-squares regression; spar- se partial least-squares regression; the electricity demand of Yunnan province; cross-validation***

I. INTRODUCTION

The demand of electricity power is increasing with the rapid development of China. Thus, it is important to analyze the factors that impact the electricity consumption as well as to predict the electricity consumption in the future. There are various researches on electricity power, here we list some of them: Alice and Lam[1]analyzed the relationship between economic development and electricity consumption based on a 30 years dataset of China. Fan and Wang[2]studied the impact of energy-saving emission reduction policy on electricity demand under low carbon economy. And Yu and Lin[3]applied the partial least squares regression to predict the electricity power demand of Shandong province.

In this paper, we manly focus on the electricity consumption demand of Yunnan Province, China. Many factors such as economic development, population, consumer price index are contained in our analysis. It's well known that the ordinary least square method (OLS) is not very suitable when it comes to high-dimensional multi-collinear data. In this paper, we employ partial least-squares regression (PLS) to deal with the electricity data. PLS is quite suitable for high-dimensional and relevant data. In addition, the number of variables can be larger than the sample size in PLS algorithm. However, the principal components of PLS is the linear combination of all the original predictors, as such, the uninformative variables are also entered the final model. Sparse partial least-squares regression (SPLS) not only inherits the advantages of PLS but also selects the key variables. In addition, the results on electricity data of Yunnan show that SPLS performs better prediction accuracy than OLS.

Sponsor: Dai Lin, female, Yunnan Kunming people, master's, professor, research interests: applied statistics.

II. SPARSE PARTIAL LEAST SQUARES REGRESSION

SPLS is proposed by Chun and Keles[4]in 2010. SPLS can be seen as the combinations of partial least squares regression (PLS) model and a sparse variable selection procedure. SPLS chooses variables by shrinking coeffients of low impact variables to zero, and reserving the major influent ones. It means that we can achieve variable selection while doing regression[5]. Namely, SPLS can automatically select important variables and eliminate the uninformative variables. SPLS is also an iteration algorithm. Firstly, we use the following formula to solve the first direction vector of sparse partial least-squares regression：

$$\max_{w}\left(\boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}\right) \quad \textbf{\textit{s.t.}} \quad \boldsymbol{w}^T\boldsymbol{w}=1 \ , \ |\boldsymbol{w}|\le\lambda \qquad (1)$$

where $\boldsymbol{M}=\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}$ ($\boldsymbol{Y}$ is a single response, and $\boldsymbol{X}=\left(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_p\right)$ is the matrix of predictors, the standardized data matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted by $\boldsymbol{E}_0=\left(\boldsymbol{E}_{01},\boldsymbol{E}_{02},\cdots,\boldsymbol{E}_{0p}\right)_{n\times p}$ and $\boldsymbol{F}_0$ respectively). And λ determi- nes the degree of sparsity. The smaller λ is the closer coeffi- ents get to zero, which leads to higher degree of sparsity. However, Jolliffe[4]pointed out that the solution of this problem is not sufficiently sparse, and the problem itself is not even a convex optimization problem. Thereupon, Chun and Kele[4]added the constraint conditions of L norm of w , then model (1) becomes:

$$\min_{w,c}\left\{-\gamma\boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}+(1-\gamma)(\boldsymbol{c}-\boldsymbol{w})^T\boldsymbol{M}(\boldsymbol{c}-\boldsymbol{w})+\lambda_1\left|\boldsymbol{c}\right|_1+\lambda_2\left|\boldsymbol{c}\right|_2^2\right\}$$
$$s.t. \quad \boldsymbol{w}^T\boldsymbol{w}=1 \qquad (2)$$

Model (2) enhances the zero attributes and assures the high correltion of $\boldsymbol{w}$ and $\boldsymbol{c}$ by adding L_1 and L_2 norm on the vector $\boldsymbol{c}$. In this formula, L_1 supports the sparsity on the vector $\boldsymbol{c}$ and L_2 can be used to solve the potential singular points. In this paper, c is normalized and treated as an estimate vector. When γ is taken as 1, it becomes the problem with the largest eigenvalues in partial least-squares regression.

A. SPLS algorithm

Theoretically, every vector can be obtained either by SIMPLS or NIPALS iteration algorithm, but we would probably lose the conjugation of vectors by doing so. Although Smith orthogonalization preserves the conjugation of vectors, the transformed vectors may not be convergent due to the loss of properties of the Krylov sequence. And it means the inaccurate evaluation[5].

Chun and Kele[4]gives the following algorithm for SPLS: Each step of NIPALS or SIMPLS algorithm is recorded to find out the active variable and the direction vectors is updated continuously. Where $\boldsymbol{A}$ is indexes of active variables, K is the number of direction vectors, and $\boldsymbol{X}_A$ is the subset of matrix $\boldsymbol{X}$, whose column index is included in $\boldsymbol{A}$ .The SPLS algorithm can be achieved by NIPALS or SIMPLS, assuming the $\boldsymbol{X}$ and $\boldsymbol{Y}$ are all standardized. And the specific procedure are as follows:

Step 1: Let $\hat{\boldsymbol{\beta}}^{\text{PLS}}=0$, $\boldsymbol{A}=\{\cdot\}$ and $k=1$. For the NIPALS algorithm, $\boldsymbol{Y}_1=\boldsymbol{Y}$, and $\boldsymbol{X}_1=\boldsymbol{X}$ for the SIMPLS algorithm.

Step 2: When $k\le K$,

(a) For the NIPALS algorithm, let $\boldsymbol{M}=\boldsymbol{X}^T\boldsymbol{Y}_1\boldsymbol{Y}_1^T\boldsymbol{X}$, and solve $\hat{\boldsymbol{w}}$ through the convex optimization formula. For the SIMPLS algorithm, let $\boldsymbol{M}=\boldsymbol{X}_1^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}_1$, and solve $\hat{\boldsymbol{w}}$ through the convex optimization formula;

(b) Updating $\boldsymbol{A}$ as $\{i:\boldsymbol{w}_i\ne 0\}\cup\{i:\hat{\boldsymbol{\beta}}^{PLS}\ne 0\}$,

(c) Using k direction vectors to obtain PLS $\hat{\boldsymbol{\beta}}^{PLS}$ in $\boldsymbol{X}_A$.

(d) Updating $\hat{\boldsymbol{\beta}}^{PLS}$ by PLS, and updating k wiht $k\leftarrow k+1$.

For the NIPALS algorithm, updating $\boldsymbol{Y}_l$ with $\boldsymbol{Y}_1\leftarrow\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{PLS}$,

For the SIMPLS algorithm, updating $\boldsymbol{X}_1$ with $\boldsymbol{X}_{1A}\leftarrow\boldsymbol{X}_A\left(\boldsymbol{I}-\boldsymbol{P}_A\left(\boldsymbol{P}_A^T\boldsymbol{P}_A\right)^{-1}\boldsymbol{P}_A^T\right)$,

where $\boldsymbol{P}_A=\boldsymbol{X}_A^T\boldsymbol{X}_A\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{X}_A^T\boldsymbol{X}_A\boldsymbol{W}_A)^{-1}$, repeat step 2 until $k=K$ [4].

B. Parameters tuning in SPLS

It is indicated in formula (2) that there are four parameters $(\gamma,\lambda_1,\lambda_2,K)$ in the sparse partial least-squares regression model, but only two of them are the key tuning parameters, namely the threshold parameter λ_1 and the number of hidden elements K . Where the parameter γ takes values from 0 to 0.5 and it will adjust both properties of concave and convex of the objective function and the similarity between $\boldsymbol{w}$ and $\boldsymbol{c}$. When consider the single-variable problem, we take $\gamma=0.5$. We integrate the constraint conditions λ_1 and λ_2 as the weight penalty factors of the objective function. And the solution of the objective function can reach convergence if only λ_2 is big enough, thus let $\lambda_2\to\infty$. Therefore it is only necessary to conduct tuning mechanism for two key parameters λ_1 and K , where K is the components number of the model. And the value of λ_1 should minimize both the approximate residual error and the nonzero component solutions. In this paper, we use cross-validation and to determine the optimal values of λ_1 and K .

Cross-validation is also referred to as cycle estimation, which always cut the data sample into the smaller data subsets and conduct analysis on arbitrary subset(we also call them the training sets), then compare with other subsets(known as validation sets or test sets) to obtain qualifications and validations. There are three main ways for cross validation. First, divide the data into two groups randomly; second, divide the data into K groups (usually divided averagely), and take a subset as the validation subset and the rest K-1 ones as the training subsets, repeat K times; third, take each sample as a validation subset separately and the rest samples as the training subsets, which is also called the leave-one-out cross validation.

Due to the small sample size n of the sparse partial least-squares regression, we tend to use the leave-one-out cross validation. Fistly, we taking out one individual(numbered i) from training set $\{\boldsymbol{X}_{train},\boldsymbol{y}_{train}\}$ in turn, then we have the remaining $n_{train}-1$ individuals as test set, repeat n_{train} times. And we obtain the optimal penalty factor λ_1 and the components number K by computing the mean squared error of prediction of cross-validation $(MSEP_{CV})$:

$$MSEP_{CV}=\frac{1}{n_{train}}\sum_{i=1}^{n_{train}}\left(y_{train,i}-\hat{y}_{train,i}\right)^2 \qquad (3)$$

Where $y_{train,i}$ and $\hat{y}_{train,i}$ are the observation and prediction of the dependent variable of the i th sample, respectively.

All the training sample $\{\boldsymbol{X}_{train},\boldsymbol{y}_{train}\}$ can be used to build the correction model and to predicte the testing sample $\{\boldsymbol{X}_{test},\boldsymbol{y}_{test}\}$ after determining the optimal penalty factor λ_1 and the components number K . And the the mean squared error of prediction of cross-validation can be computed as follows:

$$MSEP=\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}\left(y_{test,i}-\hat{y}_{test,i}\right)^2 \qquad (4)$$

C. Model prediction accuracy

Let y_i be the observation, $\hat{y}_i$ be the prediction value, and $\bar{y}$ be the average value of observation. We adopt three indexes to evaluate the prediction of the model:

(1) Multiple determination coefficient (R^2):

Let SSR be the squared sum of the residual:

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

And SSY be the squared sum of the total partial difference.

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R^2 is defined by the following equation:

$$R^2 = 1 - (SSR / SSY) \quad (5)$$

The multiple determination coefficient[6] always take values between 0 and 1.The more R^2 is close to 1, the better the regression is, which implies the significantly correlation between the independent variables and the dependent variables are. And $R^2 > 0.7$ means that the interpretation of the model is reliable. $R^2 > 0.9$ indicates the data is well fitted by the model.

(2)The error of the mean squared root:

$$RMSE = \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n \right)^{1/2} \quad (6)$$

(3)The relative prediction error:

$$RPE = \sum_{i=1}^{n} |y_i - \hat{y}_i| \Big/ \sum_{i=1}^{n} |y_i| \quad (7)$$

III. DATA

It is shown in both domestic and foreign research that there are many factors influencing the electricity consumption, such as economy, urban population and the environment factors and so on. And all these factors are highly correlated. By comprehensive comparative analysis, in this paper we determine the main factors as follows: GDP of the first industry $\boldsymbol{x}_1$ (one hundred million Yuan), GDP of the second industry $\boldsymbol{x}_2$ (one hundred million Yuan), GDP of the third industry $\boldsymbol{x}_3$ (one hundred million Yuan), the domestic gross product value $\boldsymbol{x}_4$ (Yuan/person)of per capital , fixed asset investment $\boldsymbol{x}_5$ (one hundred million Yuan), total retail sales of social consumer goods $\boldsymbol{x}_6$ (one hundred million Yuan), the agricultural population $\boldsymbol{x}_7$ (ten thousand people), non-agricultural population $\boldsymbol{x}_8$ (ten thousand people), the consumer price index $\boldsymbol{x}_9$ (CPI), the dependent variable $\boldsymbol{y}$ in this paper is the whole society power consumption(million kilowatt hour). The data is obtained from 《China statistical yearbook》 and 《Statistical yearbook of yunnan province 》 from 1999 to 2013.

Table1.The electric power consumption of Yunnan province

year	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
1999	406.87	811.90	681.05	4558	717.28	538.95	3554.8	637.6	99.70	296.70
2000	431.80	833.25	746.14	4770	697.94	583.17	3584.3	656.5	97.90	273.58
2001	444.42	868.06	825.83	5015	734.81	640.80	3609.9	677.5	99.10	320.75
2002	463.44	934.88	914.50	5366	828.65	711.25	3636.3	696.8	99.80	353.20
2003	494.60	1047.66	1013.76	5870	1021.18	782.46	3662.4	713.2	101.20	370.31
2004	593.59	1281.63	1206.69	7012	1330.60	915.31	3691.1	724.1	106	454.51
2005	661.69	1426.42	1374.62	7809	1755.30	1041.29	3720.5	729.9	101.40	557.25
2006	724.40	1705.83	1557.91	8929	2220.45	1204.75	3740.2	742.8	101.90	645.61
2007	837.35	2038.39	1896.78	10609	2798.89	1422.57	3764.3	749.7	105.90	745.52
2008	1020.56	2452.75	2218.81	12570	3526.60	1764.74	3789.4	753.6	105.70	829.44
2009	1067.60	2582.53	2519.62	13539	4527.02	2051.06	3812.8	758.2	100.40	891.19
2010	1108.38	3223.49	2892.31	15752	5528.71	2542.44	3838.3	763.3	103.73	1004.07
2011	1411.01	3780.32	3701.79	19265	6185.30	3000.14	3862.6	768.4	104.85	1204.07
2012	1654.55	4419.20	4235.72	22195	7831.10	3541.60	3885.8	773.2	102.73	1315.86
2013	1895.34	4927.82	4897.75	25083	9968.30	4004.56	3908.3	778.3	103.12	1459.81

First of all, colinearity test is carried out to examine the multi-linearity among variables. And a preliminary judgment can be conducted through correlation coeffients among variables. The correlation coefficients are presented as follows.

Table 2. The correlation coefficients

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
x_1	1.000	0.995	0.997	0.998	0.992	0.994	0.937	0.824	0.467	0.991
x_2	0.995	1.000	0.998	0.999	0.994	0.998	0.942	0.827	0.478	0.994
x_3	0.997	0.998	1.000	0.999	0.995	0.998	0.936	0.822	0.459	0.991
x_4	0.998	0.999	0.999	1.000	0.994	0.998	0.940	0.826	0.471	0.994
x_5	0.992	0.994	0.995	0.994	1.000	0.996	0.919	0.794	0.414	0.982
x_6	0.994	0.998	0.998	0.998	0.996	1.000	0.932	0.812	0.441	0.989
x_7	0.937	0.942	0.936	0.940	0.919	0.932	1.000	0.964	0.614	0.967
x_8	0.824	0.827	0.822	0.826	0.794	0.812	0.964	1.000	0.690	0.870
x_9	0.467	0.478	0.459	0.471	0.414	0.441	0.614	0.690	1.000	0.520
y	0.991	0.994	0.991	0.994	0.982	0.989	0.967	0.870	0.520	1.000

Table 2 shows that the majority of the correlation coefficients between variables $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_9$ are bigger than 0.8, which means the severe multiple colinearity.

IV. RESULTS AND DISCUSSION

A. *SPLS parameters adjustment*

The main difference between sparse partial least-squares regression and the partial least-squares regression is that before the main ingredients being extracted the former can impose the constraint conditions on weight vectors, compress the coefficients, and put the useful coefficients together to reduce the coefficients which has a little influence on dependent variables to zero, then conduct the partial least squares regression.

Because the sample size n is smaller, we take the cross validation with leaving one subset. The CV function of SPLS of R can directly return the optimal value of λ_1 and K after it gives the value ranges of the independent and dependent variables $\boldsymbol{x}$ and $\boldsymbol{y}$,and parameters λ_1 and K . Set n as the number of the sample size, $\lambda_1 \in (0,1)$, $K \in \left(1, \min\left\{p, (v-1)n / v\right\}\right)$ .Where p denotes the number of independent variables, the value v the function takes in fold n , thus we have $K \in [1,9]$ .Finally, it returns the optimal values $\lambda_1 = 0.8$, $K = 1$,namely the sparse partial least squares regression model only subtracts a principal ingredient.

The chart of the mean square error of the cross validation prediction is presented as following.

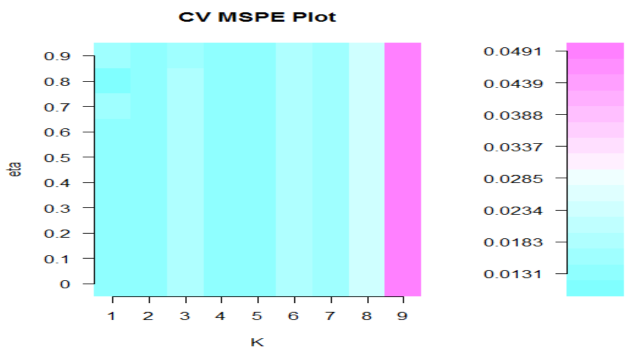

Fig.1 The chart of the mean square error of the cross validation prediction

We can find in this chart that when $\lambda_1 = 0.8$, $K = 1$, $MSEP_{CV}$ takes the minimum value 0.0131.Thus the four parameters $(\gamma, \lambda_1, \lambda_2, K)$ can be determined as $(0.5, 0.8, +\infty, 1)$.

B. SPLS modeling

When subtracting the sparse principal components, we always put the punishment on the weight vector of the original data to make some coefficients reduced to zero, then delete some unrelated variables and achieve the variables choices.

Using the coefficients $\lambda_1 = 0.80$, $K = 1$ obtained from the last section, we perform the sparse partial least squares regression to data which is standardized with the help of SPLS. And the coefficients of the dependent variables are presented as following.

Table3. Sparse regression coefficients

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
$\hat{\beta}$	0.130	0.130	0.130	0.130	0.129	0.129	0.127	0.114	0

The results from the above table show that the coefficient x_9 is 0,which implies that citizen consumption price index (CPI) x_9 has no influence on the dependent variable, thus it can be deleted automatically from the independent variables.

Finally, we get the sparse partial least squares equation of raw data as following,

$$y = -1992.114 + 0.1146x_1 + 0.0383x_2 + 0.0397x_3 + 0.0079x_4 + 0.0191x_5 + 0.0463x_6 + 0.414x_7 + 0.9202x_8$$

And the model parameters are:

$R^2 = 0.99409$, $RMSE = 29.14955$, $RPE = 0.03230$

Because the multiple determination coefficient R^2 , the mean square root error $RMSE$, and the relative prediction error RPE are small, so the model prediction efficiency is also very good.

C. Comparison with OLS and PLS

1) OLS modeling:

Using R software to construct the ordinary least-squares regressions about all the indexes of electricity power consumption of Yunnan province.

The ordinary least squares formula:

$$y = -6659.547 - 0.286x_1 + 0.365x_2 + 0.354x_3 + 0.573x_5 - 0.559x_6 + 2.515x_7 - 2.685x_8 - 4.190x_9$$

And the model parameters are:

$R^2 = 0.99005$, $RMSE = 37.82625$, $RPE = 0.03922$ 。

2) PLS modeling:

Because there are different units among all indexes, in order to eliminate the dimensional difference among all variables, first we standardize the data .In this paper we take the most common Z standardization method .After standardization for the independent variable X and the dependent variable Y , denoted by E_0 and F_0 , respectively. We subtract the first principal component t_1 , and point out that w_1 is the biggest eigenvector corresponding to the biggest eigen value of $E_0^T F_0 F_0^T E_0$:

$$w_1 = (0.354, 0.355, 0.354, 0.355, 0.351, 0.353, 0.345, 0.311, 0.186)^T$$

By $\mathbf{t}_1 = \mathbf{E}_0 \mathbf{w}_1$, the first principal component can be obtained after a simple calculation:

$$t_1 = (-3.455, -3.294, -2.907, -2.352, -2.086, -1.256, -1.185, -0.664, 0.228, 0.946, 1.083, 2.119, 3.300, 4.265, 5.438)^T$$

And since $E_0 = t_1 p_1^T + E_1$,

$$p_1 = \frac{E_0^T t_1}{\|t_1\|^2} = (0.352, 0.352, 0.352, 0.352, 0.348, 0.351, 0.347, 0.347, 0.200)^T$$

From $E_1 = E_0 - t_1 p_1^T$,we can obtain E_1 . Replace E_0 by E_1 , repeat this process and the iteration stops until the cross validity of the extracted components is less than 0.0975.

Likewise, we can get the second components:

$$w_2 = (0.133, 0.148, 0.136, 0.144, 0.163, 0.148, 0, -0.390, -0.844)^T$$

$$t_2 = (0.860, 1.277, 0.719, 0.350, -0.199, -1.712, -0.202, -0.348, -1.505, -1.227, 0.598, -0.239, -0.246, 0.810, 1.072)^T$$

The cross validities[7]of the two components are presented as following.

Table4. Cross validition

No. of scores	Q_h^2	**threshold**
1	1.000	0.0975
2	-0.1561	0.0975

It shows in table 4 that the cross validity of the first component is 1, and the one of the second component is 0.1561,which implies that the introduction of the first two components can promote a better prediction ability.

Finally, we get the partial least-squares regression equation of the raw data.

$$y = -2213.895 + 0.110x_1 + 0.038x_2 + 0.039x_3 + 0.008x_4 + 0.018x_5 + 0.046x_6 + 0.410x_7 + 0.805x_8 + 3.177x_9$$

And the model parameters are:

$R^2 = 0.99237$ ， $RMSE = 33.10996$ ， $RPE = 0.035366$

Seen from what stated above, the prediction accuracy of the partial least squares model is better than that of the ordinary least squares regression. Thus it is concluded that the partial least squares regression model can perform better than the ordinary least squares regression when the correlation coefficients matrix show the existence of severe multiple correlation among variables.

3) Comparison among OLS, PLS and SPLS :

In this paper we set the analysis of the influence factors of Yunnan province's electricity consumption as example to establish the ordinary least-squares regression and the sparse partial least-squares regression model, then compare and analysis the prediction efficiency with the help of three models.

The prediction accuracy of three methods are presented as following.

Table5 .The contrasts of $RMSE$ 、 R^2 、 RPE

Model	R^2	$RMSE$	RPE
OLS	0.99005	37.82625	0.03922
PLS	0.99237	33.10996	0.03537
SPLS	0.99409	29.14955	0.03230

Seen from the table 5,we can find that R^2 for SPLS takes the biggest value, while $RMSE$、RPE for SPLS takes the least values, which implies that SPLS presents the best fitting efficiency and the highest prediction accuracy.Thus it is concluded that the sparse partial least-squares regression model can solve the problem of correlation among variables and conduct variables choices, except that it performs higher prediction accuracy than other models. Because the data size is little, we take the method with leaving one subset to forecast, and the corresponding results are presented as the following:

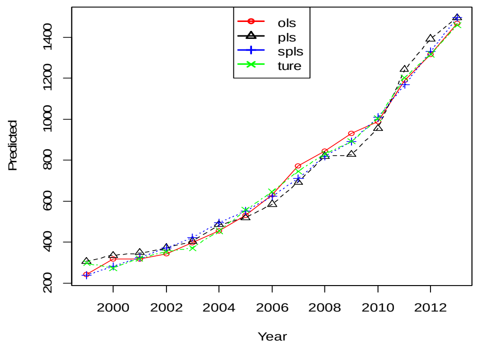

Fig.2 The prediction values from 1999 to 2013

As is shown in figure, the real values of electricity consumption of Yunnan province, and the prediction broken line of OLS, PLS and SPLS are also clearly presented in this figure. And we can find that the sparse partial least-squares regression model performs better in fitting with the real data.Then we compare the three models' coefficients, whose diagram is presented as following:

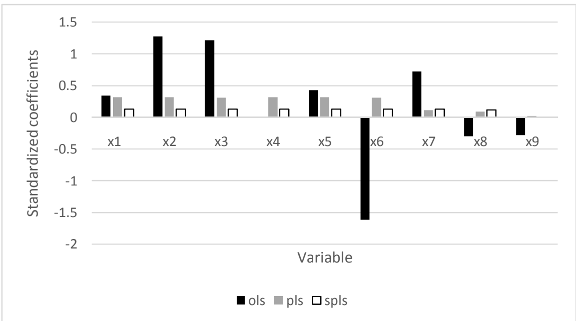

Fig.3 The bar diagram of OLS, PLS and SPLS models' coeff-icients

Seen from the figure, both x_4 and x_9 are missing. The partial reasons are that in the ordinary least-squares regressions the variance inflation factor of x_4 is far bigger than 10 and presents the biggest value, thus we delete it. While x_9 in the sparse partial least-squares regression modeling reduces automatically the independent variables coefficients which have little influence on dependent variables to zero. Here the advantage of variable selection of the sparse partial least-squares regression is highlighted .The sparse partial least squares coefficients present a good interpretation for the electivity consumption influence factors of Yunnan province. The first six independent variable coefficients shows that ,with the economy development of Yunnan province, the increase of investment in fixed assets, the growth of total retail sales of social consumer goods, the demand of the electricity of Yunnan province also raises. Compared with the partial least-squares regression model, the sparse partial least-squares regression presents a better fitting efficiency and higher, prediction accuracy.

Acknowledgment

I would like to express my warmest gratitude to Dailin, my mentor, for she instructive suggestions and valuable comments on the writing of this thesis. At the same time, paper source of fund projects is the national natural science foundation of China（11261025，11201200）,Natural science fund project in yunnan province（KKSY201207044）,Fund key projects in yunnan province department of education（2014Z011）

Conclusion

SPLS can automatically select important variables and eliminate the uninformative variables. The root mean square errors(RMSE) is used to evaluate the prediction performance and the results shows that SPLS is competitive with ordinary least squares (OLS) and partial least squares regression(PLS). In addition, several predictors such as GDP of Yunnan are chosen as key factors with SPLS algorithm.

[1] Shiu A，Lam P L. Electricity consumption and economic growth in China［J］. Energy Policy，2004(32) , pp.47－54.

[2] Fan De-cheng,Wang Shao-hua ,Zhang Wei. Electric power demand forecasting research in China under low carbon economy target［J］. The grid technology，2012，36(7) , pp.19－25. (in Chinese)

[3] Yu Qing-song, Lin Cheng. Based on partial least-squares regression analysis of the Shan dong electric power demand forecasting[J]. Resources and environment in arid areas，2015，29（2）, pp.14-20. (in Chinese)

[4] Hyonho Chun and Sündüz Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection.[J] Journal of the Royal Statisyical Society ISSN，2010，72(1) , pp.3-25.

[5] Wang feng. Based on sparse partial least-square algorithm study biological omics data fusion algorithm [D] . Ji lin : Jilin university，2012. (in Chinese)

[6] He Xiao-qun , Liou Wen-ching. Application of regression analysis[M]. Beijing: Ren min university of China publishing house,2011. (in Chinese)

[7] Wang Pan. Partial least squares regression method in the application of the commercial housing sales price analysis[J]. Statistics and decision, 2014,(07), pp.84-86. (in Chinese)

[8] Wang Hui-wen. Partial least squares regression method and its application [M]. Beijing: National defence industry press,1999. (in Chinese)

[9] Tang Ying-chun. Weighted fusion method and comparison of the sparse partial least squares method [D] . Ji lin:The northeast normal university.，2010. (in Chinese)

[10] Liu Dao-hua . Based on the RBF neural network model of the orthogonal least squares[J] Journal of xinyang normal university: natural science,2013，26(3) , pp.428-431. (in Chinese).