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-dimentional 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; sparse 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-dimension 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.

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 coefficients 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:

$$\max_w (w^T M w) \quad s.t. \quad w^T w = 1, \quad |w| \leq \lambda$$  \hspace{1cm} (1)

where $M = X^TYX$ (Y is a single response, and $X = [x_1, x_2, \ldots, x_p]$) is the matrix of predictors, the standardized data matrix of X and Y are denoted by $E_x = [E_{x_1}, E_{x_2}, \ldots, E_{x_p}]$ and $E_y$ respectively. And $\lambda$ determines the degree of sparsity. The smaller $\lambda$ is, the closer coefficients 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 $L_0$ norm of $w$, then model (1) becomes:

$$\min_w \left\{ -\gamma w^T M w + (1-\gamma)(c-w)^T M (c-w) + \lambda_1 |w| + \lambda_2 |c|^2 \right\} \quad s.t. \quad w^T w = 1$$  \hspace{1cm} (2)

Model (2) enhances the zero attributes and assures the high correlation of $w$ and $c$ by adding $L_1$ and $L_2$ norm on the vector $c$. In this formula, $L_1$ supports the sparsity on the vector $c$ 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 $\gamma$ 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 $A$ is indexes of active variables, $K$ is the number of direction vectors, and $X_i$ is the subset of matrix $X$, whose column index is included in $A$. The SPLS algorithm can be achieved by NIPALS or SIMPLS, assuming the $X$ and $Y$ are all standardized. And the specific procedure are as follows:

Step 1: Let $\hat{P}_{PLS} = 0$, $A = \{\}$ and $k = 1$. For the NIPALS algorithm, $y_i = y_i$, and $x_i = x_i$ for the SIMPLS algorithm.

Step 2: When $k \leq K$ ,

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

(b) Updating $A$ as $\{i : w_i \neq 0\} \cup \{i : \hat{P}_{PLS} \neq 0\}$ ,

(c) Using $k$ direction vectors to obtain PLS $\hat{P}_{PLS}$ in $X_i$.

(d) Updating $\hat{P}_{PLS}$ by PLS, and updating $k$ with $k \leftarrow k + 1$.

For the NIPALS algorithm, updating $y_i$ with $y_i \leftarrow y_i - X_i \hat{P}_{PLS}$ .

For the SIMPLS algorithm, updating $X_i$ with $X_i \leftarrow X_i \left(1 - P_i (P_i' P_i)^{-1} P_i'\right)$ , where $P_i = X_i' X_i W_i (W_i' X_i X_i W_i)^{-1}$ , 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 $\lambda_1$ and the number of hidden elements $K$. Where the parameter $\gamma$ 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 $w$ and $c$. When consider the single-variable problem, we take $\gamma = 0.5$ . We integrate the constraint conditions $\lambda_1$ and $\lambda_2$ as the weight penalty factors of the objective function. And the solution of the objective function can reach convergence if only $\lambda_2$ is big enough, thus let $\lambda_2 \rightarrow \infty$ . Therefore it is only necessary to conduct tuning mechanism for two key parameters $\lambda_1$ and $K$, where $K$ is the components number of the model. And the value of $\lambda_1$ 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 $\lambda_1$ 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(know 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. Firstly, we taking out one individual(numbered i) from training set $\{X_{train}, y_{train}\}$ 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 $\lambda_1$ and the components number $K$ by computing the mean squared error of prediction of cross-validation ($MSEP_{cv}$):

$$MSEP_{cv} = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} (y_{train,i} - \hat{y}_{train,i})^2$$

(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 $\{X_{train}, y_{train}\}$ can be used to build the correction model and to predict the testing sample $\{X_{test}, y_{test}\}$ after determining the optimal penalty factor $\lambda_1$ and the components number $K$. And the the mean squared error of prediction of cross-validation can be computed as follows:

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

(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:

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

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:

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

The multiple determination coefficient[6] always takes 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:

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

(3) The relative prediction error:

$$RPE = \sum_{i=1}^{n} |y_i - \hat{y}_i| / \sum_{i=1}^{n} |y_i|$$

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 $x_1$ (one hundred million Yuan), GDP of the second industry $x_2$ (one hundred million Yuan), GDP of the third industry $x_3$ (one hundred million Yuan), the domestic gross product value $x_4$ (Yuan/person) of per capital, fixed asset investment $x_5$ (one hundred million Yuan), total retail sales of social consumer goods $x_6$ (one hundred million Yuan), the agricultural population $x_7$ (ten thousand people), non-agricultural population $x_8$ (ten thousand people), the consumer price index $x_9$ (CPI), the dependent variable $y$ 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.

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>x8</th>
<th>x9</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>406.87</td>
<td>811.99</td>
<td>681.05</td>
<td>4558</td>
<td>717.28</td>
<td>538.95</td>
<td>3554.8</td>
<td>637.6</td>
<td>99.70</td>
</tr>
<tr>
<td>2000</td>
<td>431.80</td>
<td>833.25</td>
<td>746.14</td>
<td>4770</td>
<td>697.84</td>
<td>583.17</td>
<td>3584.3</td>
<td>656.5</td>
<td>97.90</td>
</tr>
<tr>
<td>2001</td>
<td>444.42</td>
<td>868.06</td>
<td>825.83</td>
<td>5015</td>
<td>734.81</td>
<td>640.80</td>
<td>3609.9</td>
<td>677.5</td>
<td>99.10</td>
</tr>
<tr>
<td>2002</td>
<td>463.44</td>
<td>934.88</td>
<td>914.50</td>
<td>5366</td>
<td>711.25</td>
<td>6363.6</td>
<td>699.8</td>
<td>98.80</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>494.60</td>
<td>1047.66</td>
<td>1013.76</td>
<td>5870</td>
<td>722.46</td>
<td>6462.4</td>
<td>4713.2</td>
<td>101.20</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>539.59</td>
<td>1281.63</td>
<td>1206.69</td>
<td>7012</td>
<td>7153.69</td>
<td>9153.1</td>
<td>5691.1</td>
<td>7241.4</td>
<td>106</td>
</tr>
<tr>
<td>2005</td>
<td>661.69</td>
<td>1426.42</td>
<td>1374.62</td>
<td>7809</td>
<td>749.7</td>
<td>1755.30</td>
<td>1041.29</td>
<td>3702.5</td>
<td>577.25</td>
</tr>
<tr>
<td>2006</td>
<td>724.40</td>
<td>1705.83</td>
<td>1557.91</td>
<td>8929</td>
<td>1204.75</td>
<td>1755.30</td>
<td>1041.29</td>
<td>3702.5</td>
<td>577.25</td>
</tr>
<tr>
<td>2007</td>
<td>780.04</td>
<td>1981.38</td>
<td>1794.82</td>
<td>10699</td>
<td>1204.75</td>
<td>1755.30</td>
<td>1041.29</td>
<td>3702.5</td>
<td>577.25</td>
</tr>
<tr>
<td>2008</td>
<td>807.36</td>
<td>2083.19</td>
<td>1974.78</td>
<td>10699</td>
<td>1204.75</td>
<td>1755.30</td>
<td>1041.29</td>
<td>3702.5</td>
<td>577.25</td>
</tr>
<tr>
<td>2009</td>
<td>829.44</td>
<td>2185.72</td>
<td>21570</td>
<td>3526.60</td>
<td>1764.74</td>
<td>3789.4</td>
<td>105.70</td>
<td>10792.4</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>900.60</td>
<td>2528.52</td>
<td>2519.62</td>
<td>13539</td>
<td>4527.02</td>
<td>1051.06</td>
<td>3812.8</td>
<td>7582.8</td>
<td>100.40</td>
</tr>
<tr>
<td>2011</td>
<td>902.38</td>
<td>2922.49</td>
<td>2982.31</td>
<td>15752</td>
<td>5528.71</td>
<td>1782.44</td>
<td>3883.8</td>
<td>7633.1</td>
<td>103.73</td>
</tr>
<tr>
<td>2012</td>
<td>911.41</td>
<td>3780.32</td>
<td>3701.79</td>
<td>192656185.30</td>
<td>1948.20</td>
<td>4323.72</td>
<td>221957831.10</td>
<td>3541.60</td>
<td>3858.5</td>
</tr>
<tr>
<td>2013</td>
<td>13895.34</td>
<td>4927.82</td>
<td>4897.75</td>
<td>25083</td>
<td>9666.30</td>
<td>4004.56</td>
<td>53908.3</td>
<td>7783.7</td>
<td>103.12</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
<th>x7</th>
<th>x8</th>
<th>x9</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>0.995</td>
<td>0.979</td>
<td>0.998</td>
<td>0.992</td>
<td>0.994</td>
<td>0.937</td>
<td>0.824</td>
<td>0.467</td>
<td>0.991</td>
</tr>
<tr>
<td>0.995</td>
<td>1.000</td>
<td>0.998</td>
<td>0.999</td>
<td>0.994</td>
<td>0.998</td>
<td>0.942</td>
<td>0.827</td>
<td>0.478</td>
<td>0.994</td>
</tr>
<tr>
<td>0.997</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.995</td>
<td>0.998</td>
<td>0.936</td>
<td>0.822</td>
<td>0.459</td>
<td>0.991</td>
</tr>
<tr>
<td>0.996</td>
<td>1.000</td>
<td>0.994</td>
<td>1.000</td>
<td>0.994</td>
<td>0.996</td>
<td>0.919</td>
<td>0.794</td>
<td>0.414</td>
<td>0.982</td>
</tr>
<tr>
<td>0.996</td>
<td>1.000</td>
<td>0.998</td>
<td>0.998</td>
<td>0.996</td>
<td>1.000</td>
<td>0.932</td>
<td>0.812</td>
<td>0.441</td>
<td>0.989</td>
</tr>
<tr>
<td>0.998</td>
<td>0.994</td>
<td>0.998</td>
<td>0.998</td>
<td>0.996</td>
<td>1.000</td>
<td>0.932</td>
<td>0.812</td>
<td>0.441</td>
<td>0.989</td>
</tr>
<tr>
<td>0.932</td>
<td>0.942</td>
<td>0.936</td>
<td>0.940</td>
<td>0.941</td>
<td>0.919</td>
<td>0.919</td>
<td>0.894</td>
<td>0.414</td>
<td>0.982</td>
</tr>
<tr>
<td>0.812</td>
<td>0.827</td>
<td>0.822</td>
<td>0.826</td>
<td>0.794</td>
<td>0.812</td>
<td>0.812</td>
<td>0.894</td>
<td>0.414</td>
<td>0.989</td>
</tr>
<tr>
<td>0.967</td>
<td>0.870</td>
<td>0.520</td>
<td>1.000</td>
<td>0.690</td>
<td>0.870</td>
<td>0.870</td>
<td>0.520</td>
<td>1.000</td>
<td>0.520</td>
</tr>
</tbody>
</table>

Table 2 shows that the majority of the correlation coefficients between variables $x_1, x_2, \ldots, x_9$ 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 $\lambda_i$ and $K$ after it gives the value ranges of the independent and dependent variables $x$ and $y$ ,and parameters $\lambda_i$ and $K$. Set $n$ as the number of the sample size, $K$ is the optimal number of the principal ingredients.

![Fig.1 The chart of the mean square error of the cross validation prediction](image-url)
We can find in this chart that when $\lambda = 0.8$, $K = 1$, $MSEP_2$ 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 $\lambda = 0.80$, $K = 1$ 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.

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>0.130</td>
<td>0.130</td>
<td>0.130</td>
<td>0.129</td>
<td>0.129</td>
<td>0.127</td>
<td>0.114</td>
<td>0.010</td>
</tr>
</tbody>
</table>

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,

$$y = -1992.114 + 0.114x_1 + 0.0383x_1 + 0.0397x_1 + 0.0079x_1 + 0.0191x_1 + 0.0465x_1 + 0.414x_1 + 0.9202x_1$$

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:

$$y = -6659.547 - 0.286x_1 + 0.365x_1 + 0.354x_1 + 0.573x_1 - 0.559x_1 + 2.515x_1 - 2.685x_1 - 4.190x_1$$

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 $E_x$ and $E_y$, 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_x^T F_y F_y^T F_y^T .

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

By $t_1 = E_x w_1$, the first principal component can be obtained after a simple calculation:

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

since $E_y = t_1 p_{y1}^T + E_y$, we can obtain $E_y$. Replace $E_y$ by $E_y$, 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:

$$w_2 = (0.133, 0.148, 0.136, 0.144, 0.163, 0.148, 0.0, 0.390, -0.844)^T$$

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

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

<table>
<thead>
<tr>
<th>No. of scores</th>
<th>$Q^2$</th>
<th>threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>0.0975</td>
</tr>
<tr>
<td>2</td>
<td>-0.1561</td>
<td>0.0975</td>
</tr>
</tbody>
</table>

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:

$$y = -2213.895 + 0.110x_1 + 0.038x_1 + 0.039x_1 + 0.008x_1 + 0.018x_1 + 0.046x_1 + 0.410x_1 + 0.805x_1 + 3.177x_1$$

And the model parameters are:

$$R^2 = 0.99237, \ RMSE = 33.10996, \ RPE = 0.03536$$

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 analyze the prediction efficiency with the help of three models.

The prediction accuracy of three methods are presented as following.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>RPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>0.99005</td>
<td>37.82625</td>
<td>0.03922</td>
</tr>
<tr>
<td>PLS</td>
<td>0.99237</td>
<td>33.10996</td>
<td>0.03537</td>
</tr>
<tr>
<td>SPLS</td>
<td>0.99409</td>
<td>29.14955</td>
<td>0.03230</td>
</tr>
</tbody>
</table>

Seen from the table 5, we can find that $R^2$ for SPLS
takes the biggest value, while \( \text{RMSE} \) and \( \text{RPE} \) 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:

\section*{Conclusion}

SPLS can automatically select important variables and eliminate the uninformative variables. The root mean square error (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.

\section*{Acknowledgment}

I would like to express my warmest gratitude to Dailin, my mentor, for her 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 (KKSYY201207044), Fund key projects in yunnan province department of education (2014Z011)

\begin{thebibliography}{9}
\bibitem{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)
\end{thebibliography}