Sensors and Materials, Vol. 32, No. 10 (2020) 3463–3478
MYU Tokyo

S & M 2351

Partial Least Squares Optimization Method and Path Analysis Integration for Chinese Medicine Data

Tianci Li,[1] Wangping Xiong,[1*] Jianqiang Du,[1] Bin Nie,[1]
Jigen Luo,[1] Yanyun Yang,[1] and Chih-Cheng Chen[2**]

[1]School of Computer Science, Jiangxi University of Traditional Chinese Medicine,
Nanchang, Jiangxi 330004, China
[2]School of Information and Engineering College, Jimei University, Xiamen, Fujian 361021, China

(Received April 22, 2020; accepted August 18, 2020)

Keywords: Chinese medicine dosage, partial least squares (PLS), path analysis, variable selection

Partial least squares (PLS) is widely used in multivariate statistical analysis, but linear and nonlinear model variable selections are based on the selection of principal components. It does not involve the interactions of variables and predictors, which may adversely affect prediction accuracy. In this study, we design a tailor built temperature control system to monitor and control temperature settings during experiments on traditional Chinese medicine (TCM). We combine results from path analysis and the variables' covariance and correlation matrix, and propose a PLS optimization method that integrates path analysis (PLS-PA). To verify the validity of PLS-PA, we use the measured coefficients and residuals as evaluation indicators. We test the performance of PLS-PA using two TCM dose datasets and one dataset from the University of California, Irvine (UCI). The three experimental results demonstrate that the measured coefficients from the traditional PLS and PLS-PA methods increase by 11.8, 4.7, and 8.5%, which suggest the validity of our experiment. We conclude that PLS-PA can optimize the screening of variables and improve the PLS regression analysis of TCM experimental data without hampering model accuracy.

1. Introduction

The process of decocting traditional Chinese medicine (TCM) involves careful consideration of the concentration of reactants, the ratio of ingredients, reaction time, decocting temperature, and pressure, which are necessary for an accurate reaction. Among these factors, accurately controlling the temperature affects the reaction rate, which is a crucial factor in the TCM experimental process. The effect of temperature on TCM experiments is complex and mainly includes the following points:

(1) The reaction rate of a TCM experiment increases exponentially with temperature.
(2) In a certain temperature range, the experimental reaction speed increases with temperature, but when the temperature exceeds a certain level, an increase in temperature decreases the reaction speed.
(3) As the temperature rises, the reaction rate of the experiment decreases.

*Corresponding author: e-mail: 20030730@jxutcmedu.cn
**Corresponding author: e-mail: 201761000018@jmu.edu.cn

https://doi.org/10.18494/SAM.2020.2931

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

In addition, minor differences in boiling temperature will also affect the content of chemical components. For example, Zhu *et al.*(1) proved that 115 °C is the best boiling temperature for a Dachengqi decoction, and increasing or decreasing the temperature will decrease the content of some chemical components. Therefore, it is necessary to monitor the temperature in real time during a TCM experiment. Experiments at a constant fixed and optimal boiling temperature can avoid differences in chemical content due to temperature.

In multivariate regression, when the number of independent variables is much larger than that of sample points, the least squares method does not solve the problem very well. Partial least squares (PLS), however, integrates the basic functions of principal component, canonical correlation, and linear regression analyses,(2) and effectively combines data analysis results used for prediction with nonmodal data cognitive analysis results. Data analysis helps find the functional relationship between dependent and independent variables in order for a prediction model to be established.(3) PLS simplifies the data structure through data analysis and finds the relationship between variables. Compared with linear regression and ridge regression analyses, PLS is thus a better regression method for multiple independent variables linked to multiple dependent variables.

The composition and mechanisms of TCM are complicated.(4) Most of the clinical and experimental data consist of noisy information, multiple correlations of variables, and nonlinearity,(5) which are acceptable conditions for the application of PLS regression. However, the number of variables in TCM experimental data tends to be large, and even some observation data are difficult to manage. Some independent variables may have little or no effect on some dependent variables. If the PLS regression model incorporates these unrelated variables, the amount of computational calculations will increase and the model will be inaccurate. If too many independent variables are selected, then results will be collinear. Conversely, if some important variables are missed, the regression results will be affected, and sometimes unpredictable parameter estimates will be generated. Therefore, when using TCM experimental data to establish a PLS multiple regression model,(6) screening the variables in the regression model is necessary to effectively improve its accuracy.

2. Related Research

At present, the quality of temperature detection systems is generally reflected in the instrument's level of sophistication, the temperature's range of measurement, the measurement precision, and the instrument's power consumption.(7) However, in China, the accuracy of temperature detection is inadequate, because the majority of control systems are single-parameter single-loop systems controlled by a single-chip microcomputer. Multiparameter comprehensive control systems do not exist, resulting in a large gap in the sophistication of temperature control systems. Also, for the same type of TCM experiment, temperature detection accuracy can be very different depending on the laboratories used. Despite the rapid development of computing, some laboratories have yet to fully realize the importance of accurate temperature detection and control during a TCM experiment, which may severely affect experimental results.

Temperature control systems in China, such as those studied by Wang *et al.*,[8] used a single-chip microcomputer for temperature control. This method can ensure relatively highly accurate temperature control, but temperature transmission is subject to delay. Li *et al.*'s control system[9] incorporated the proportional-integral-derivative (PID) segmentation theory to make the system respond quickly to temperature changes, but the entire system lacked stability. Kong *et al.*'s remote temperature control system[10] greatly reduced labor costs, but its fault maintenance and operation costs were relatively high. The advantages and disadvantages of the system developed in this study and other systems are shown in Table 1.

Typical variable screening methods such as stepwise regression, subset extraction, and optimal subset variable screening have their own advantages and disadvantages. Stepwise regression and subset extraction methods are very commonly used, but the random errors in the variable screening process are neglected, so it is difficult to systematically study their theoretical properties. Analysis with the optimal subset variable screening rule lacks stability. The well-known Akaike information criterion (AIC) and Bayesian information criterion (BIC)[11] are selection criteria based on the Kullback–Leibler information distance and minimized Bayes posterior probability. These methods statistically depend on the likelihood of the model, and in general, the study of likelihood functions requires knowledge of the type of distribution, with only some parameters unknown. These conditions increase the difficulty in using the aforementioned variable screening methods.

Some variable screening methods apply a punishment function in their statistical analysis. The central idea of this type of method is to replace the minimized loss function based on the penalization of least squares, minimizing the sum of the loss and penalty functions. The classical penalty functions include the least absolute shrinkage and selection operator (Lasso),[12] smoothly clipped absolute deviation (SCAD) penalty,[13] Adaptive Lasso,[14] and Elastic Net.[15]

However, with regard to variable screening methods that utilize a punishment function, the calculation is difficult and sometimes impossible for $p >> n$. In response to this problem, Cai and Lv[16] proposed a Dantzig screening method. This variable screening method involves L1 normalization under the condition that the design matrix satisfies the uniform uncertainty principle (UUP). This method has some desirable mathematical properties, such as if the problem can be transformed into a linear specification problem, then it is easy to solve. However, when the number of dimensions increases and the UUP condition is not easily satisfied, there is no guarantee that the correct model can be selected.

Table 1
Comparison of proposed temperature control system with other systems.

	Comparisons			
Name of system	Data deviation of temperature sampling	Stability of temperature control	Time lag of temperature feedback	Cost-effectiveness
GTCS[8]	Y	N	N	Y
PID-PTCS[9]	Y	N	Y	N
RIMSUCR[10]	N	Y	Y	N
This paper	Y	Y	Y	Y

To solve the limitation of the Dantzig screening method, Song *et al.*(17) and Fan and Lv(18) proposed a sure independent screening (SIS) method, which can reduce the number of dimensions from p to m, where $m < n$. Yuan and Lin(19) and Kim *et al.*(20) carried out a similar study of grouping variables and applying a punishment function. However, in practical problems, the following phenomena may occur: some unimportant variables may be preferentially selected because of collinearity with other important variables, and other variables may not be highly correlated with dependent variables individually, but some independent variables may be strongly correlated with dependent variables when combined with independent variables. Therefore, Fan and Lv(21) proposed an iterative SIS (ISIS) method to overcome these problems.

In addition, there are many methods for the selection of linear model variables, but they are generally based on the model error, which has a normal distribution, and are established by the least squares, penalized least squares, wider Lq,(22) or penalty Lq method. In particular, for a large p and a small n, that is, the dimension p is larger than the number of samples, n, the least squares, penalized least squares, wider Lq, and penalty Lq methods perform inaccurately and are slow, and sometimes the results obtained are not completely uniform owing to the use of different penalty functions. These inconveniences make it very difficult to apply screening methods. When linear variable selection cannot meet the requirements of nonlinear variable selection, it must be combined with other nonlinear regression methods,(23) such as artificial neural networks, nuclear methods,(24) support vector machines,(25) and PLS. Wang(26) used the kernel function as the transformation basis function to study the nonlinear structural features of data by PLS regression based on a kernel function transformation.

Path analysis is a statistical method that involves a path graph and multiple linear regressions.(27) It is capable of visualizing the relationship between independent and dependent variables. It is also capable of calculating the direct effect of each causal factor on outcome factors and the indirect effect on the output variable through a path coefficient and the calculation of a residual path coefficient. At the same time, a path map can be used to visually indicate relationships that are difficult to express in multivariate analysis and also help indicate the importance of the different variables in relation to output variables. The development of path analysis as a linear regression model helps overcome the limitations of linear regression models.

In this paper, to tackle the complexity of the TCM dose–effect treatment problem, we use TCM experiments and information from the literature to construct a complete path map with the independent and dependent variables as nodes and the direct and indirect path coefficients as weights. Through the weight analysis of the directed weight graph, different comprehensive weights of various paths are obtained, and the independent variable point groups with large direct and indirect effects on the dependent variable are selected according to weight. At the same time, the principal components of PLS and the principal component of the dependent variable path coefficient are calculated. By retaining a PLS suitable for modeling with a sample size less than the number of variables, a new variable screening method for fusion path analysis is proposed, which uses the PLS regression model for optimization.

3. Methods

Path analysis is separated into a direct path coefficient (the direct effect of one independent variable on the dependent variable) and an indirect path coefficient (the sum of the indirect effects of an independent variable on a dependent variable by affecting other independent variables) on the basis of multiple regression with correlation coefficients.

For general multivariate linear regression, we set independent variables X_1, X_2, ..., X_n and the dependent variable Y.

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_n X_n \tag{1}$$

$$\overline{Y} = B_0 + B_1 \overline{X}_1 + B_2 \overline{X}_2 + \cdots + B_n \overline{X}_n \tag{2}$$

Subtracting Eq. (2) from Eq. (1) gives

$$Y - \overline{Y} = B_1(X_1 - \overline{X}_1) + B_2(X_2 - \overline{X}_2) + \cdots + B_n(X_n - \overline{X}_n)\,. \tag{3}$$

Dividing both sides of Eq. (3) by the standard deviation δ_y of Y at the same time gives

$$\begin{aligned}(Y - \overline{Y}) / \delta_y &= B_1(X_2 - \overline{X}_2) / \delta_y + B_2(X_2 - \overline{X}_2) / \delta_y + \cdots + B_n(X_n - \overline{X}_n) / \delta_y \\ &= B_1 \frac{\delta_{X_1}}{\delta_y} \frac{X_1 - \overline{X}_1}{\delta_{X_1}} + B_2 \frac{\delta_{X_2}}{\delta_y} \frac{X_2 - \overline{X}_2}{\delta_{X_2}} + \cdots + B_n \frac{\delta_{X_n}}{\delta_y} \frac{X_n - \overline{X}_n}{\delta_{X_n}}\end{aligned} \tag{4}$$

The models of linear regression coefficients of respective variables in Eq. (4) are obtained by using the least squares method. From this equation, the decomposition equation of each simple correlation coefficient can be obtained by carrying out quantity transformation:

$$\begin{cases} P_{1Y} + r_{12}P_{2Y} + r_{13}P_{3Y} + \cdots + r_{1n}P_{nY} = r_{1Y} \\ r_{21}P_{1Y} + P_{2Y} + r_{23}P_{3Y} + \cdots + r_{2n}P_{nY} = r_{2Y} \\ \qquad \vdots \\ r_{n1}P_{1Y} + r_{n2}P_{2Y} + r_{n3}P_{3Y} + \cdots + P_{nY} = r_{nY} \end{cases} . \tag{5}$$

Equation (5) is the basic path analysis model, during which r_{ij} is the simple correlation coefficient of X_i and X_j. Moreover, r_{iY} is the simple correlation coefficient of X_i and Y. P_{iY} is the partial correlation coefficient between X_i and Y after standardization, showing that X_i has a direct effect on Y. $r_{ij}P_{jY}$ is the indirect path, indicating the indirect effect of X_i on Y by affecting X_j.

Through software linear regression, the resulting standard coefficient is the size coefficient that we require, which is then multiplied by the correlation coefficient to obtain indirect path coefficients.

Before the path analysis is carried out for PLS regression, an auxiliary independent variable is selected. That is, the path analysis is used to calculate the effect of each independent variable

X_j on the explanatory variable Y, and the information irrelevant to the explanatory variable is eliminated. Regression modeling is then performed using methods such as canonical correlation and multiple linear analyses, and cross-validation to verify the predictive power of the model. The entire experimental process is shown in Fig. 1.

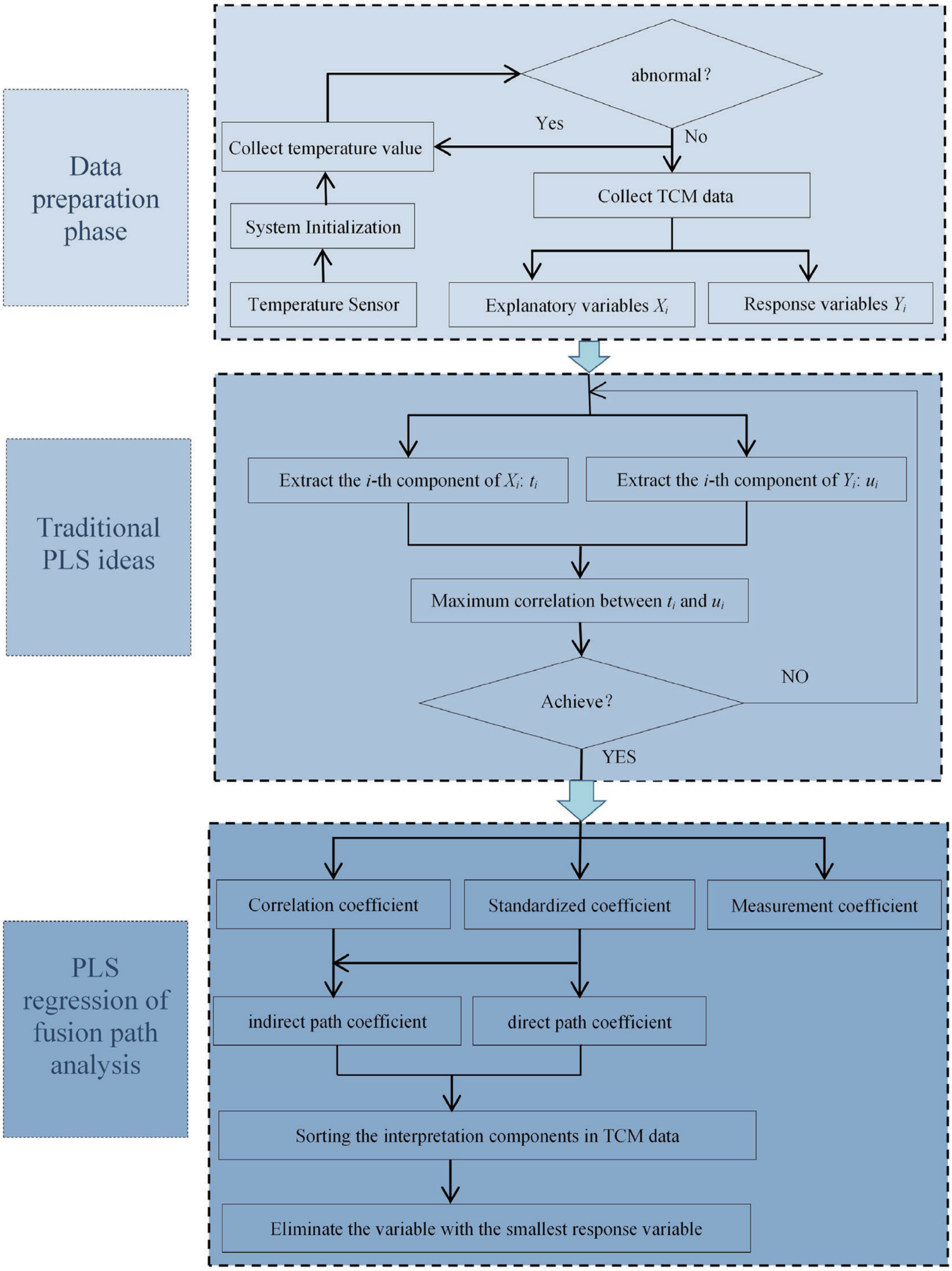

Data preparation phase
abnormal?
Yes
No
Collect temperature value
System Initialization
Temperature Sensor
Collect TCM data
Explanatory variables X_i
Response variables Y_i
Traditional PLS ideas
Extract the i-th component of X_i: t_i
Extract the i-th component of Y_i: u_i
Maximum correlation between t_i and u_i
Achieve?
NO
YES
PLS regression of fusion path analysis
Correlation coefficient
Standardized coefficient
Measurement coefficient
indirect path coefficient
direct path coefficient
Sorting the interpretation components in TCM data
Eliminate the variable with the smallest response variable

Fig. 1. (Color online) Flowchart of experiment.

The experiment includes the following four steps to prove the effectiveness of the PLS-PA method.

(1) Establish a PLS regression model for sample data and obtain the standard and correlation coefficients through linear regression. Calculate the coefficient of determination from the regression model.

(2) Obtain the direct path coefficient from the standard regression coefficient and combine the correlation coefficients of sample data for analysis and filtering. Remove the variable with the most adverse effect on sample regression.

(3) Process the sample data obtained in the previous step to establish a PLS regression model and once again calculate the coefficient of determination of the model.

(4) Compare the measurement coefficient before and after optimization for any improvement.

The algorithm flow is described in Algorithm 1.

Algorithm 1: PLS optimization algorithm combined with path analysis

```
 Input: Original sample dataset ($Z$)
Output: PLS-PA equation.
Begin: Use the path analysis method for the dataset $Z = (X,Y)$
    Calculate the path coefficient matrix of $Z$
  IF $X$ and $Y$ are related
        Remove independent variables with an adverse effect on regression
        (1) Standardize dataset $Z$
        (2) Calculate $\omega_i, \upsilon_i$: the maximum eigenvalues and eigenvectors of $X_{i-1}^T Y_{i-1} Y_{i-1}^T X_{i-1}$, $Y_{i-1}^T X_{i-1} X_{i-1}^T Y_{i-1}$
        (3) Using the feature vector $\omega_i, \upsilon_i$ calculate the score vector: $t_i = X_{i-1}\omega_i$, $\mu_i = Y_{i-1}\upsilon_i$
        (4) Calculate the regression coefficient vector: $P_i = (X_{i-1}^T t_i)/ \| t_i \|^2$, $r_i = (Y_{i-1}^T t_i)/ \| t_i \|^2$
   End
  PLS regression is used to calculate the multiple linear regression equations of $Y$ to $X$
 End
```

4. Experimental Results and Discussion

4.1 Temperature control system

Our system is a self-developed temperature control system, and each boiling process in an experiment is carried out at a constant and optimal temperature. After decocting, TCM data are collected at 60 ± 2 °C. The main modules of the system include the medicine information input and medicine information output. The system monitors the temperature in real time and alerts experimenters if the temperature is abnormal. The modules used are shown in Table 2.

Table 2
Modules and devices of the system.

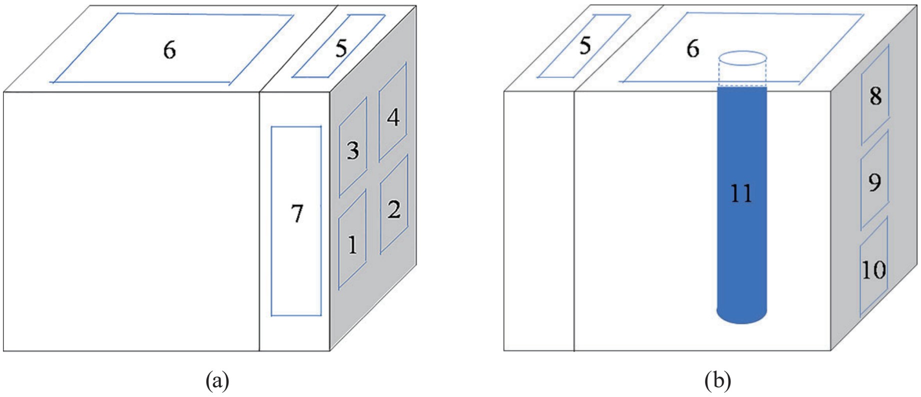
Module	Device
Display module	Display screen
Data storage module	Data storage chip
Reminder module	Speaker
Text-to-speech module	TTS voice module
Temperature measurement module	Infrared thermometer
Controller module	MSP430 Chip

The temperature control module includes a digital-to-analog conversion module (1), a refrigeration module (2), and a high-current drive module (7). The digital-to-analog conversion module (1) is connected to a processing module (4); the high-current drive module (7) is connected to a digital-to-analog conversion module (1), and the cooling module (2) is connected to a high-current drive module (7). The specific module design is shown in Figs. 2 and 3 below.

The temperature is intelligently controlled within suitable ranges for a TCM experiment through the main control chip, temperature detection real-time display, temperature control, and function keyboard modules. We try to reduce the number of environmental factors that may change the experimental results. The module that debugs the received data is shown in Fig. 4.

4.2 Experimental platform and sample

We mainly use the temperature control system to monitor the temperature, collect the TCM data within the same temperature range, and then test the performance of the fitting of PLS optimization method integrating path analysis (PLS-PA) during a TCM experiment. The algorithm is programmed by MATLAB. Datasets A and B in the experiment are from the Key Laboratory of Modern Chinese Medicine Preparations, Ministry of Education, and the third dataset comes from the Concrete Compressive Strength on the UCI dataset.



(a) (b)

Fig. 2. (Color online) Structure of intelligent temperature control system.

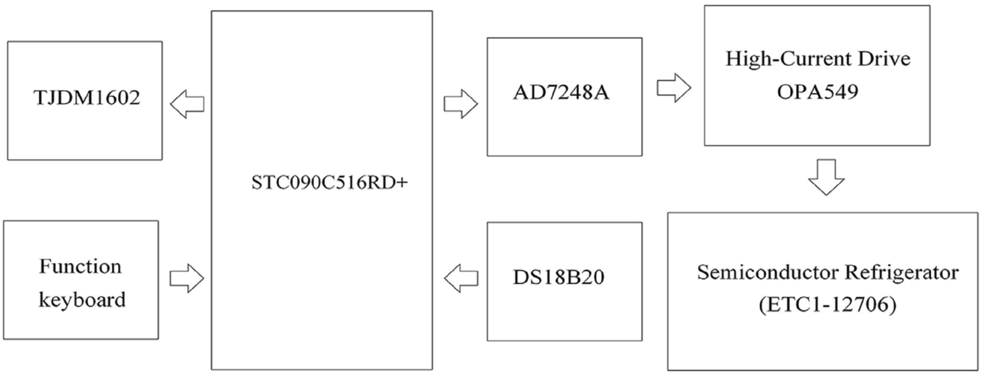

TJDM1602
Function keyboard
STC090C516RD+
AD7248A
High-Current Drive OPA549
DS18B20
Semiconductor Refrigerator (ETC1-12706)

Fig. 3. Intelligent temperature control system module.

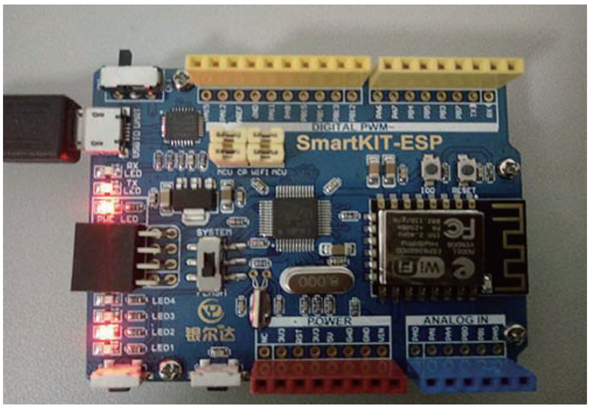


Fig. 4. (Color online) Module for debugging received data.

The first experimental dataset A is used to study the changes in the physiological index of superoxide dismutase (SOD) under different dosages of rhubarb, magnolia, citrus aurantium, and mirabilite. The independent variables are rhubarb, magnolia, citrus, and mirabilite (set to x_1–x_4), and the dependent variable is SOD (set to y). There are nine groups of experimental samples. The data table is shown in Table 3.

The second experimental dataset B is used to study the changes in the active components in plasma in the treatment of intestinal obstruction with different proportions of a Dachengqi decoction. There are nine independent variables in the dataset and one dependent variable (small intestinal circumference). There are twelve groups of data in total. The data table is shown in Table 4.

In the third experimental dataset, there are eight independent variables that comprise the compressive strength of one concrete and dependent variable. The total number of samples is 1030. A detailed description of this dataset can be found at http://archive.ics.uci.edu/ml/.

To verify the improved performance of the PLS-PA method, the above experimental data are compared with those obtained by the traditional PLS and PLS–variable importance in projection (PLS-VIP) value optimization methods. In our comparison, all raw data are randomly divided in a ratio of 7:3, with the former used as experimental samples and the latter used as test data.

4.3 Evaluation index

(1) Residual e_i

For a good fitting equation, the sum of the residuals should be as small as possible. The smaller the residual, the closer the fitted equation is to the observed values, that is, the greater the ability of the fitting equation to interpret y. The following equation is used in the calculation of the residual:

$$e_i = y_i - \hat{y}_i \,. \tag{6}$$

Table 3
Experimental data table for dose–effect relationship.

Rhubarb	Magnolia	Citrus aurantium	Mirabilite	SOD
x_1	x_2	x_3	x_4	y
41.93	17.02	90.96	5.36	0.05
9.16	69.10	4.28	72.72	0.03
23.50	100.70	15.53	15.53	0.04
2.93	44.62	89.75	17.95	0.05
32.03	26.89	26.77	69.57	0.06
15.96	3.93	82.72	52.65	0.06
69.82	50.56	25.20	9.69	0.06
96.01	22.29	14.36	22.57	0.13
53.95	8.82	15.42	77.07	0.22

Table 4
Experimental data table for different dose ratios.

Emodin	Rhein	Chrysophanol	…	Hesperetin	Small intestinal circumference
x_1	x_2	x_3	…	x_9	y
3.9733	162.3330	12.5160	…	0.6733	2.5714
1.7683	32.7030	4.4567	…	0.2600	2.4286
5.1900	129.4540	17.5520	…	0.7360	2.6143
3.2733	80.5460	9.3883	…	0.9533	2.4000
⋮	⋮	⋮	…	⋮	⋮
5.0400	154.5400	11.3125	…	0.1075	2.5143
6.6460	119.8540	5.9560	…	2.8360	2.4857
13.7780	142.0500	18.5980	…	0.1920	2.5714

(2) Measured coefficient R^2

R^2 indicates the number of interpretable mutations as a percentage of the total number of variations from the data, thus indicating the line fitting performance from regression. It also indicates the degree of correlation of the dependent variable y with the fitted variable. From this viewpoint, the greater the correlation between the fitted variable and the original variable y, the better the fit of the fitted line. The calculation formula for the measurement coefficient R^2 is given below:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{7}$$

where SST represents the total variation of the sum of squares of the original data y_i; SSR is the explanatory variation of the sum of squares with a degree of freedom of 1.

4.4 Experimental results and analysis

Using MATLAB, the direct path coefficients of all the independent variables in relation to the dependent variables are calculated, and we obtain all the tables of path coefficients. See Tables 5–7. In the tables, the underlined values are the direct path coefficients, and the variables are indirect path coefficients, that is, the indirect effect of X_i on Y by affecting X_j.

Table 5
Path coefficient table of experimental dataset A.

	x_1	x_2	x_3	x_4
x_1	0.4014	0.1212	0.0800	0.0684
x_2	−0.1933	−0.4078	0.0863	0.0698
x_3	−0.1538	0.1687	−0.2086	0.1064
x_4	−0.1000	0.1033	0.0806	0.2754

Table 6
Path coefficient table of experimental dataset B.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	−0.3134	−0.1450	−0.0324	−0.0819	0.0339	−0.0018	−0.0578	−0.0018	0.0020
x_2	−0.1399	−0.3249	−0.0294	−0.1734	0.0462	−0.0032	−0.0888	0.0109	0.0056
x_3	−0.2138	−0.2013	−0.0475	−0.1344	0.0245	−0.0009	−0.0676	−0.0006	−0.0007
x_4	−0.0854	−0.1874	−0.0212	0.3006	0.0526	−0.0036	−0.0265	0.0038	0.0023
x_5	−0.1178	−0.1665	−0.0129	−0.1751	0.0902	−0.0062	−0.0862	0.0075	0.0058
x_6	−0.0651	−0.1177	−0.0050	0.1228	0.0638	−0.0088	−0.0481	0.0044	0.0067
x_7	−0.1308	−0.2081	−0.0232	−0.0575	0.0561	−0.0031	−0.1386	0.0077	0.0024
x_8	0.0289	−0.1831	0.0016	−0.0589	0.0350	−0.0020	−0.0547	0.0194	0.0073
x_9	−0.0492	−0.1433	0.0027	−0.0559	0.0418	0.0047	−0.0259	0.0113	0.0126

Table 7
Path coefficient table of UCI dataset.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0.4010	−0.1103	−0.1594	−0.0327	0.0372	−0.0438	−0.0893	0.0329
x_2	−0.0481	0.1749	−0.0566	0.0188	0.0076	−0.0497	−0.0493	−0.0077
x_3	0.0134	0.0109	−0.0336	0.0086	−0.0127	0.0003	−0.0027	0.0052
x_4	0.0219	−0.0288	0.0690	−0.2685	0.1765	0.0489	0.1210	−0.0745
x_5	0.0242	0.0113	0.0986	−0.1718	0.2613	−0.0696	0.0581	−0.0503
x_6	0.0090	0.0233	0.0008	0.0150	0.0219	−0.0821	0.0146	0.0002
x_7	0.0455	0.0575	−0.0161	0.0920	−0.0454	0.0364	−0.2042	0.0319
x_8	0.0335	−0.0181	−0.0631	0.1135	−0.0788	−0.0012	−0.0638	0.4088

As can be seen from Table 5, for experimental dataset A, the effect of x_3 on y is minimal, and the other independent variables x_i (i = 1, 2, 3, 4) have a small effect on y by affecting x_3, so we delete the variable x_3. Similarly, for experimental dataset B, x_6 has the least effect on y, so we delete this variable. For the UCI dataset, the excluded variable is x_3.

At the same time, to verify the feasibility and effectiveness of our PLS-PA method, we analyze the importance of the variables through the PLS-VIP method and compare the results with those of the PLS-VIP method. The results are shown in Figs. 5–7. It can be observed that the variables deleted by the two variable screening methods are not the same. For example, for experimental dataset A, the variable deleted by PLS-PA is x_3, and the variable deleted by the PLS-VIP method is x_4. This difference also shows that because of the strong correlation and redundancy between the variables in Chinese medicine data, there is a strong mutual effect among variables.

Two methods are used to separate the three new samples with the currently removed independent variables. Using MATLAB, we established a PLS regression model. The residuals (e_i) and measured coefficients (R^2) of the regression model are obtained. The comparison of the residuals obtained before and after optimization is shown in Table 8, and the measurement

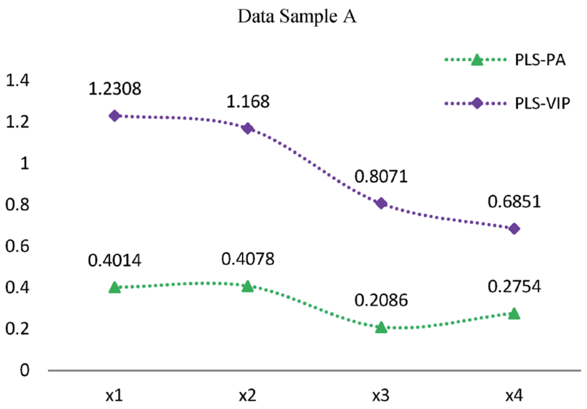


Fig. 5. (Color online) Comparison of importance of variables for dataset A.

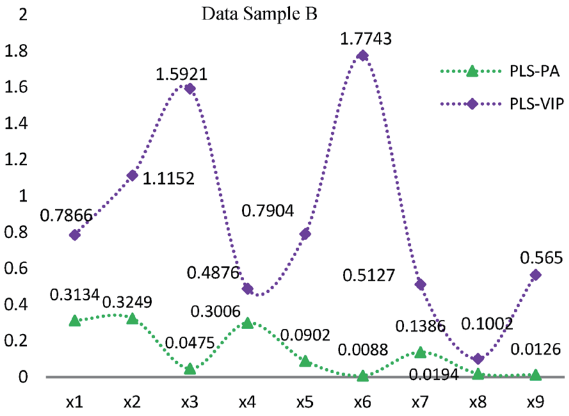


Fig. 6. (Color online) Comparison of importance of variables for dataset B.

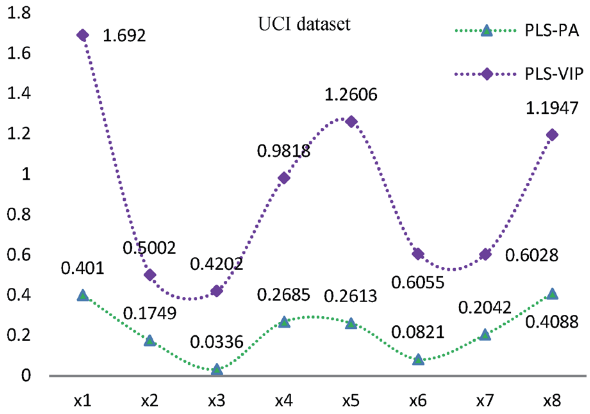


Fig. 7. (Color online) Comparison of importance of variables for UCI dataset.

Table 8
Comparison of residuals obtained before and after optimization.

	PLS	PLS-PA
Dataset A	1.7568	1.0436
Dataset B	0.8501	0.3605
UCI dataset	0.1138	0.0065

coefficients of the conventional PLS, PLS-PA, and PLS-VIP methods are compared. The results are shown in Table 9. To more intuitively express the structure of the data in the table, the corresponding line graphs are drawn in Figs. 8 and 9.

(1) Residual analysis

From Table 8, it is concluded that for these three experimental datasets, the PLS-PA method significantly reduces the ratio of residuals compared with the traditional PLS method. The explanatory ability of the PLS-PA model is enhanced and the fitting effect is improved.

Table 9
Measurement coefficients obtained from three methods.

	PLS		PLS-PA		PLS-VIP	
	Variable to be deleted	R^2	Variable to be deleted	R^2	Variable to be deleted	R^2
Dataset A	—	0.5842	x_3	0.6533	x_4	0.6533
Dataset B	—	0.8368	x_6	0.8759	x_4, x_8	0.8296
UCI dataset	—	0.5824	x_3	0.6334	x_3	0.5833

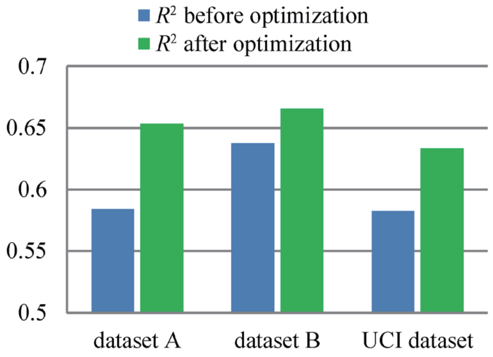


Fig. 8. (Color online) Comparison of R^2 values obtained before and after optimization.

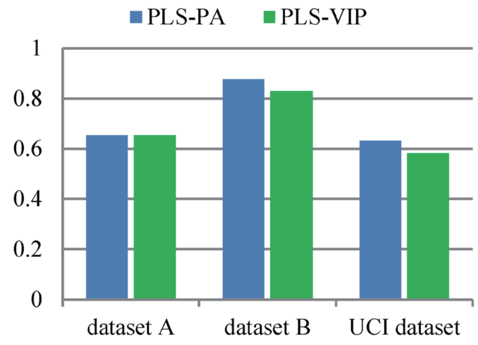


Fig. 9. (Color online) Comparison of R^2 values of two algorithms.

(2) Comparative analysis of R^2 between PLS and PLS-PA methods

From Table 9 and Fig. 8, after performing path analysis to improve the PLS regression, the determination coefficient of dataset A increased by 0.0691, compared with that of the PLS method, which increased by 11.8%. Dataset B is used to remove the variable x_3 by path analysis and R^2 is increased by 0.0391, which is an increase of 4.7%. R^2 of the UCI dataset is increased by 0.051, which is an increase of 8.5%. The results of these three datasets show that the PLS regression method after fusion path analysis is more effective than the traditional PLS method.

(3) Comparative analysis of R^2 values of two improved methods

The improvement of the PLS-VIP method compared with PLS is shown in Table 9. The comparison of the data showed that the PLS-VIP method has no clear effect on the TCM data and may even have a negative effect. According to Fig. 9, the PLS-PA method proposed in this paper is superior to the PLS-VIP method, except that dataset A is excluded from the same independent variable. The method shows that the PLS-PA method has clear advantages in the field of TCM data.

5. Conclusion

In this study, an independent temperature control system was developed to overcome the lack of temperature monitoring in TCM experiments. The system is mainly used for temperature detection during TCM experiments and temperature control during data acquisition. It can help a user avoid differences in data collection due to temperature variations in laboratories and enhance the credibility of experiments. With the aim of dealing with the complexity of variables in TCM data, a PLS optimization method incorporating path analysis was proposed to enhance the predictive ability of the model.

PLS regression can measure the importance of features based on regression coefficients, and then the top *k* features are selected to find multiple correlations between variables. However, the obtained feature subset does not usually fit well with the data. When the relationship between variables is complicated or some explanatory variables indirectly affect the response variables through other explanatory variables, there will often be a major impact on modeling. The instability of the model, which affects the parameter estimation of the model, will also increase the error of the model.

PLS includes a VIP value variable screening analysis method that has an effect on the screening of independent variables, but when the independent variables are of strong importance in relation to the dependent variables and are highly correlated between other independent variables, the effect of using VIP to filter variables is not optimal. Thus, PLS itself needs a variable screening method to solve the problem.

In light of these problems, we chose to perform path analysis to analyze the linear relationship between explanatory and response variables, thereby assisting the PLS model and increasing accuracy. Through theoretical and experimental analyses of the dose–effect TCM data, we obtain a direct path coefficient and a direct path diameter coefficient as an important measurement to determine the selection of variables. The experimental results for three datasets show that the path analysis performed to filter the independent variable can improve the regression coefficient when applying PLS. The following conclusions can be drawn:

(1) Temperature monitoring and control during an experiment are helpful for improving the quality of data and reducing noisy experimental results.

(2) The improved regression model can eliminate interfering elements in sample TCM data; thus, the data structure can be simplified and accuracy can be improved. We see that the effect of this improved regression model is better than that of PLS.

(3) When there is a strong correlation between redundant features in a data sample, the value of the direct path coefficient can effectively indicate the importance of an explanatory variable to a response variable.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 61762051 and 61562045), the Jiangxi Province Fund Major Projects (No. 20171ACE50021), and the Jiangxi Province Science and Technology Plan Projects (No. 20171BBG70108).

References

1 H. Zhu, L. Tang, and B. Jiang: J. Baotou Med. **39** (2015) 208. https://doi.org/CNKI:SUN:BTYI.0.2015-04-012

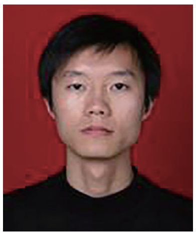
2 H. W. Wang: Partial Least Squares Regression Method and Its Application (National Defence Industry Press, Beijing, China, 1999).

3 Z. P. Zeng, J. Q. Du, R. Y. Yu, B. Nie, and F. Yu: J. Comput. Appl. **34** (2017) 87. https://doi.org/10.3969/j.issn.1001-3695.2017.01.017

4 M. Jiang and A. P. Lyu: Chin. J. Chin. Mater. Med. **39** (2014) 2149. https://doi.org/10.4268/cjcmm20141141
5 H. Y. Yu and C. G. Cheng: Chin. J. Exp. Tradit. Med. Form. **19** (2013) 343. https://doi.org/10.11653/syfj2013140343
6 G. D. Yang: Liaoning J. Tradit. Chin. Med. **44** (2017) 558. https://doi.org/10.7666/d.Y2426136
7 X. H. Zou: Inform. Tech. **10** (2005) 119. https://doi.org/10.13274/j.cnki.hdzj.2005.10.035
8 X. Y. Wang, Z. W. Zhou, and T. Wu: Autom. Instrum. **3** (2013) 63. https://doi.org/10.3969/j.issn.1001-9227.2013.03.028
9 B. W. Li, Y. Z. Zhang, and W. J. Chen: ORDNANCE Indust. Automat. **9** (2011) 30. https://doi.org/10.3969/j.issn.1006-1576.2011.09.023
10 X. Kong, J. K. Liu, and S. M. Yan: Electron. Technol. **10** (2011) 38. https://doi.org/10.3969/j.issn.1000-0755.2011.10.011
11 H. Akaike: IEEE T. Automat. Contr. **19** (1974) 716. https://doi.org/10.1109/TAC.1974.1100705
12 L. A. Lac and S. Hossain: J. Stat. Comput. Sim. **88** (2018) 3230. https://doi.org/10.1080/00949655.2018.1511713
13 H. Wang: J. Shanxi Datong Univ. **35** (2019) 38. https://doi.org/10.7666/d.Y2339347
14 B. M. Nguelifack and I. Kemajou-Brown: J. Stat. Comput. Sim. **89** (2019) 2051. https://doi.org/10.1080/00949655.2019.1607346
15 S. Mohebbi, E. Pamukcu, and H. Bozdogan: J. Stat. Comput. Sim. **89** (2019) 1060. https://doi.org/10.1080/00949655.2019.1576683
16 T. T. Cai and J. Lv: Ann. Stat. **35** (2007) 2313. https://doi.org/10.1214/009053607000000442
17 R. Q. Song, Y. Z. Zhu, and X. J. Wang: Stat. Decis. **2** (2019) 13. https://doi.org/10.13546/j.cnki.tjyjc.2019.02.003
18 J. Q. Fan and J. C. Lv: J. R. Stat. Soc. B. **70** (2008) 849. https://doi.org/10.1111/j.1467-9868.2008.00674.x
19 M. Yuan and Y. Lin: J. R. Stat. Soc. B. **69** (2007) 143. https://doi.org/10.1111/j.1467-9868.2007.00581.x
20 Y. Kim, J. Kim, and Y. Kim: Stat. Sinica **16** (2006) 375. https://doi.org/10.1007/s11135-005-8095-2
21 J. Fan and J. Lv: Stat. Sinica **20** (2010) 101. https://doi.org/10.1051/epjconf/20100402005
22 T. Z. Wu, Y. Chu, and Y. Yu: J. Appl. Math. Decis. Sci. **1** (2007) 13. https://doi.org/10.1155/2007/24053
23 E. W. Bai, K. Li, W. X. Zhao, B. Q. Mu, and W. X. Zheng: Sci. Chin. Math. **46** (2016) 1383. https://doi.org/10.1360/N012016-00002
24 R. Fezai, K. Abodayeh, M. Mansouri, H. Nounou, and M. Nounou: ICCAD (2019) 1. https://doi.org/10.1109/ICCAD46983.2019.9037905
25 D. Chen, M. Yan, Q. Li, L. Yu, Y. Jin, and K. Xu: Nanotechnol. Precis. Eng. **13** (2015) 226. https://doi.org/10.13494/j.npe.20140125
26 H. W. Wang: Linear and Nonlinear Methods for Partial Least Squares Regression (National Defense Industry Press, Beijing, 2006).
27 J. Meng: 2009 Chinese Control and Decision Conf. (CCDC, 2009) 2179.

About the Authors

Tianci Li received his B.S. degree from Hubei University of Traditional Chinese Medicine, China, in 2018 and is currently studying for an M.S. degree at Jiangxi University of Traditional Chinese Medicine. His main research directions include Chinese medicine information science, machine learning, and medical data mining. (670918214@qq.com)



Wangping Xiong received his B.S. degree from Central South University, China, in 2003 and his M.S. degree from Yunnan University, China, in 2009. He is now an associate professor in the School of Computer Science, Jiangxi University of Traditional Chinese Medicine, China. His research interests include medical data mining, hospital informatization, and medical natural language processing. (20030730@jxutcm.edu.cn)

Jianqiang Du received his B.S. degree from the Department of Precision Instruments, Tsinghua University, China, in 1992 and his M.S. degree from Huazhong University of Science and Technology CAD Center, China, in 1994. His Ph.D. degree in computer sciences was awarded by Huazhong University of Science and Technology in 2008. He is currently the vice president of Jiangxi University of Traditional Chinese Medicine. His research interests include Chinese medicine data mining, Chinese medicine informatics, medical artificial intelligence, medical data mining, and medical natural language processing. (jianqiang_du@163.com)

Bin Nie received his M.S. degree in computer science and technology from Nanchang University, China. He is now an associate professor in the School of Computer Science, Jiangxi University of Traditional Chinese Medicine, China. His research interests include data mining, machine learning, artificial intelligence, and Chinese medicine informatics. (ncunb@163.com)

Jigen Luo received his B.S. degree from Jiangxi University of Traditional Chinese Medicine, China, in 2016 and his M.S. degree in computer science and technology from Jiangxi University of Traditional Chinese Medicine in 2019. He is currently teaching at Jiangxi University of Traditional Chinese Medicine. His research interests include Chinese medicine data mining and medical natural language processing. (iluojg@163.com)

Yanyun Yang received her B.S. degree from Jiangxi University of Traditional Chinese Medicine, China, in 2018 and is currently studying for her M.S. degree at Jiangxi University of Traditional Chinese Medicine. Her main research directions include Chinese medicine information science, machine learning, and natural language processing. (1441910806@qq.com)

Chih-Cheng Chen has been a professor at Jimei University, China, since 2017. He became a member of IEEE in 2011 and a senior member in 2016. He earned his M.S. and Ph.D. degrees from the Department of Mechatronics Engineering, National Changhua University of Education. He has been studying RFID application in various fields of industry. His research interests include AIoT technology and RFID applications. (201761000018@jmu.edu.cn)