

# Discussion on Importance of Variable Selection in PLS1 Modeling

Jie Wang<sup>1</sup> Huiwen Wang<sup>1</sup> Gilbert Saporta<sup>2</sup>

<sup>1</sup> School of economic and management, Beihang University, Beijing, China, e-mail

[wangjiecom@sina.com](mailto:wangjiecom@sina.com) [wanghw@vip.sina.com](mailto:wanghw@vip.sina.com)

<sup>2</sup> Conservatoire National des Arts et Métiers, Paris, France, e-mail

[saporta@cnam.fr](mailto:saporta@cnam.fr)

**Abstract:** The multicollinearity in the independent variable sets is harmful to OLS Regression. PLS Regression, invented by Wold, brought an important breakthrough to modeling under the condition of multicollinearity. PLSR enables modeling when the multicollinearity in independent variable sets exists or the sample size is smaller than the number of independent variables, and all independent variables can be involved in the regression model. Applying PLS Regression, some researchers believe that variable selection and multicollinearity could be neglected when using PLS Regression. And in some practical cases, tens, or even hundreds, of variables are involved in the regression model. This paper indicated that the multicollinearity in independent variable sets in PLS1 can obviously affect the deriving of components and the regression parameters. Thus it is necessary to select independent variables carefully before building PLS1 models; otherwise, the regression model can still lead to unexplainable results.

**Keywords:** PLS1 Model; multicollinearity; variable selection

Multiple linear regression is the most widely used technique for quantitative analysis in many science fields. But for a long time, the multicollinearity in independent variable sets has been the main restriction of this method in application. In practice, to comprehensive describe and analyze a system, without missing some crucial features of the system as far as possible, analysts tend to select more variables. That often result in serious multiple correlation in independent variable sets. If the classic multiple regression method become invalid.

To make a breakthrough to the constraint, a lot of researches have been done. One means is to delete variables that are relevant less important. There are also some other techniques to overcome multicollinearity. For example, Ridge Regression Analysis (Hoerl, 1962), and the Principal Component Regression.

Wold and Albano et al (1983) proposed the Partial Least Squares Regression. PLSR enables modeling when the multicollinearity of independent variable sets exists or the sample size is smaller than the number of independent variables. Some researchers believe that variable selection and multicollinearity could be neglected when using PLS Regression. This paper indicates that the multicollinearity in independent variable sets has significant influence on deriving of components and the regression parameter estimation. Therefore, it is still necessary to select independent variables carefully before building PLS1 models.

## 1 A brief profiles of PLS1 regression

PLS model with only one variable is called PLS1 model. Note  $y \in \mathbb{R}^n$  as dependent variable, and  $x_j \in \mathbb{R}^n$ ,  $j=1,2,\dots,p$  as the independent variable set. For the sake of convenience, without losing generality, assume all these variables are standardized. Following are the steps of PLS1 regression.

(1) Derive component  $t_1$  from independent variable set  $X_0 = (x_1, x_2, \dots, x_p)$ , with  $t_1 = X_0 \cdot w_1$ ,  $\|w_1\| = 1$

$$w_1 = \frac{X_0^T Y}{\|X_0^T Y\|} = \frac{1}{\sqrt{\sum_{i=1}^p r^2(x_i, y)}} \begin{pmatrix} r(x_1, y) \\ r(x_2, y) \\ \dots \\ r(x_p, y) \end{pmatrix} \quad (1)$$

$$t_1 = X_0 w_1 = \frac{1}{\sqrt{\sum_{i=1}^p r^2(x_i, y)}} \left[ \sum_{i=1}^p r(x_i, y) \cdot X_{0i} \right] \quad (2)$$

After deriving  $t_1$ , regress  $X_0$  and  $y$  on  $t_1$ . Note  $X_1$  and  $y_1$  as fitting errors,

$$X_0 = t_1 p_1^T + X_1, \quad \text{with } p_1 = \frac{X_0^T t_1}{\|t_1\|^2}; \quad y_0 = t_1 r_1 + y_1, \quad \text{with } r_1 = \frac{y_0^T t_1}{\|t_1\|^2}$$

(2) Substitute  $X_0$  and  $y$  with  $X_1$  and  $y_1$ , and repeat step (1),

$$w_2 = \frac{X_1^T Y_1}{\|X_1^T Y_1\|} = \frac{1}{\sqrt{\sum_{i=1}^p \text{cov}^2(x_{1i}, y_1)}} \begin{pmatrix} \text{cov}(x_{11}, y_1) \\ \text{cov}(x_{12}, y_1) \\ \dots \\ \text{cov}(x_{1p}, y_1) \end{pmatrix}$$

$$t_2 = E_1 w_2$$

Regress  $X_1$  and  $y_1$  on  $t_2$ . Note  $X_2$  and  $y_2$  as fitting errors

$$X_1 = t_2 p_2^T + X_2, \quad \text{with } p_2 = \frac{X_1^T t_2}{\|t_2\|^2}; \quad y_1 = t_2 r_2 + y_2, \quad \text{with } r_2 = \frac{y_1^T t_2}{\|t_2\|^2}$$

Step(3), (4), ..... may be deduced by analogy. Stop deriving with the method of Cross Validation. The regression model then will be:

$$y_0 = r_1 t_1 + r_2 t_2 + \dots + r_m t_m + y_m \quad (3)$$

An important advantage of modeling (3) is that all components  $t_1, t_2, \dots, t_m$  are orthogonal vector group. Transfer model (3) into a formula of the original variables  $x_1, x_2, \dots, x_p$

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (4)$$

$$\beta_j = \sum_{h=1}^m r_h \cdot w_{hj}^*, \quad \text{with } w_h^* = \prod_{j=1}^{h-1} (I - w_j p_j^T) \cdot w_h \quad (5)$$

Formula(5) shows, if  $x_j$  contributes more to  $t_h$  (or,  $w_{hj}^*$ ) and  $t_h$  contributes more in interpreting  $y$  (that is, the greater  $F$ ), the regression coefficient of  $x_j$  will be greater. This shows the mechanism of PLS1 is reasonable. However, when applying PLS1 model, the physical meaning of the model still may be unreasonable.

## 2 Effect of multicollinearity on derivation of components

This section investigates how the multicollinearity affects the deriving of components by means of simulation analysis. First, generate two linear-independent random variables  $x_1, x_2$ , and design the model as  $y=5x_1+x_2$ . After datum standardization, the model becomes:

$$y = 0.9478 \times x_1 + 0.2164 x_2$$

To illustrate how the multicollinearity affect PLS1 model, we successively import 8 variables  $x_j = x_2 + \varepsilon_j$ ,  $j = 3, 4, \dots, 10$ , which are nearly completely correlated with  $x_2$ . ( $\varepsilon$  follows Uniform Distribution in  $(-0.1\bar{x}_2, 0.1\bar{x}_2)$ ). With the introduction of each  $x_j$ , SIMCA-P is used to calculate PLS1 regression models.

Two components are derived automatically in each modeling process and the final models are as follows:

$$y = 0.9483 \times x_1 + 0.1123 \times x_2 + 0.1058 \times x_3$$

$$y = 0.9481 \times x_1 + 0.0755 \times x_2 + 0.0692 \times x_3 + 0.0740 \times x_4$$

$$y = 0.9480 \times x_1 + 0.0566 \times x_2 + 0.0506 \times x_3 + 0.0554 \times x_4 + 0.0561 \times x_5$$

...

$$y = 0.9479 \times x_1 + 0.0285 \times x_2 + 0.0228 \times x_3 + 0.0277 \times x_4 + 0.0281 \times x_5 + \dots + 0.0234 \times x_{10}$$

An investigation on the correlation coefficients shows how the multicollinearity affects the deriving of

components. Table 1 gives the correlation coefficients of  $t_1$  to  $x_1, x_2$ .

Table 1. Correlation coefficients

variable set:	$x_1, x_2$	$x_1, x_2, x_3$	$x_1 \dots x_4$	$x_1 \dots x_5$	$x_1 \dots x_6$	$x_1 \dots x_7$	...	...	$x_1 \dots x_{10}$
$r(t_1, x_1)$	0.95	0.85	0.74	0.65	0.58	0.52	...	...	0.41
$r(t_1, x_2)$	0.44	0.64	0.76	0.84	0.88	0.91	...	...	0.96

Table 1 shows that the introduction of new variables affects the correlation coefficients between  $t_1$  and  $x_j, j=1,2$ , significantly. With the gradually introduction of  $x_3, x_4, x_5, \dots$ , the correlation coefficient of  $t_1$  and  $x_1$  decrease steadily, and that of  $t_1$  and  $x_2$  rise. After the introduction of  $x_3, x_4$ , the correlation coefficient of  $t_1$  and  $x_2$  becomes greater than that of  $t_1$  and  $x_1$ . When the independent variable set is  $x_1, x_2, x_3, \dots, x_{10}$ , the correlation coefficient of  $t_1$  and  $x_1$  dive to 0.41, and that of  $t_1$  and  $x_2$  rockets to 0.96. The reason is that the added variables, which are highly correlated with  $x_2$ , affect the deriving of PLS components. And component  $t_1$  is moved away from  $x_1$ , which are higher related to  $y$ . This case indicates that, if variables are selected arbitrarily, multicollinearity may shift  $t_1$  towards the vectors which are highly correlated.

When PLS1 is applied, an arbitrary introduction of more variables may decrease  $R^2$ . In Figure 1, the three curves are fitting errors of regression models with the following three independent variable sets,  $x_1, x_2$ ;  $x_1, x_2, x_3$  and  $x_1, x_2, x_3, x_4, x_5$ . The output of SIMCA-P shows that, with the gradually introduction of  $x_3, x_4, x_5$ , the fitting error increase. When the independent variable set is  $x_1, x_2$ , SIMCA-P derive one component automatically, with  $R^2=0.9769$ . When the independent variable set is  $x_1, x_2, x_3$ , two components are derived,  $R^2=0.9766$ . When the independent variable set is  $x_1, x_2, x_3, x_4, x_5$ , two components are derived, with  $R^2=0.9578$ .

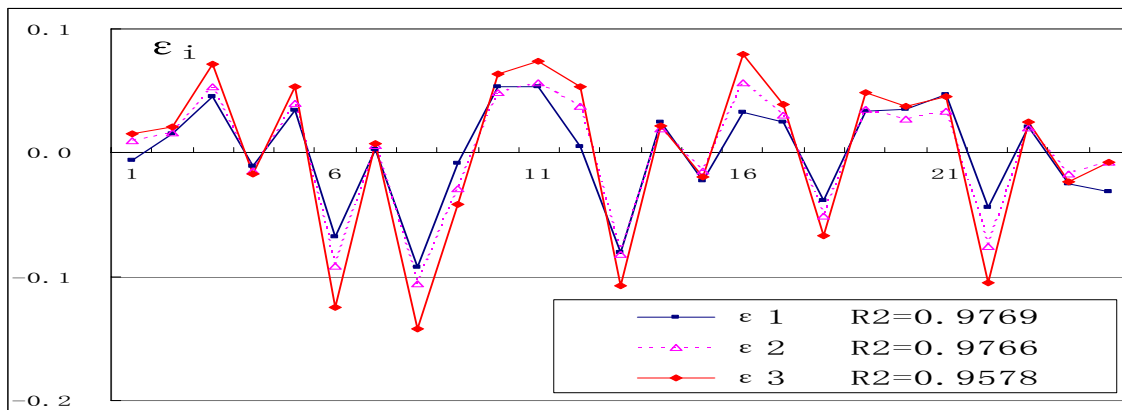


Figure 1. The increase of fitting error and decrease of  $R^2$

### 3 Effect of multicollinearity on regression coefficients

In this section, we investigate how the multicollinearity affects the regression coefficients. Generate two unrelated random variables  $x_1, x_2$ . The model is designed as  $y=x_1+2x_2$ . After standardization, the model becomes:

$$y = 0.3709x_1 + 0.922x_2 \quad (6)$$

Gradually, introduce independent variables  $x_j = x_2 + \varepsilon_j, j = 3, 4, \dots, 9$ , where  $\varepsilon$  follows a uniform distribution in  $(-0.1 \times \bar{x}_2, 0.1 \times \bar{x}_2)$ . These variables are obviously highly correlated with  $x_2$ . Calculate regression models on every  $x_j$ .

Table 2. Output of SIMCA-P

variable set	$x_1, x_2$	$x_1, x_2, x_3$	$x_1 \dots x_4$	$x_1 \dots x_5$	$x_1 \dots x_6$	$x_1 \dots x_7$	$x_1 \dots x_8$	$x_1 \dots x_9$
$b_1$	0.37579	0.36901	0.36775	0.36695	0.36763	0.36741	0.36684	0.36681
$b_2$	0.92254	0.46308	0.30831	0.23087	0.18482	0.15388	0.13162	0.11507

$b_3$	—	0.46249	0.30783	0.23043	0.18442	0.15351	0.13124	0.1147
$b_4$	—	—	0.3095	0.23208	0.18606	0.15512	0.13288	0.11634
$b_5$	—	—	—	0.2323	0.18628	0.15536	0.1331	0.11658
$b_6$	—	—	—	—	0.18412	0.15319	0.13092	0.11439
$b_7$	—	—	—	—	—	0.15464	0.13239	0.11584
$b_8$	—	—	—	—	—	—	0.13355	0.11703
$b_9$	—	—	—	—	—	—	—	0.11577
No. of components	1	2	2	2	2	2	2	2
$R^2$	0.9891	0.9995	0.9995	0.9994	0.9995	0.9995	0.9995	0.9995

Table 2 shows that, when the independent variable set is  $x_1, x_2$ , the regression coefficient of  $x_2$  is greater than that of  $x_1$ . The regression model coincides with formula (6). However, it seems that  $x_2$  is less important than  $x_1$  in the model since the introduction of  $x_4$ . This will mislead the understanding of the regression model.

#### 4 Summary

PLS is an advanced regression modeling method, to a certain extent, PLSR overcomes the affect of multicollinearity. However, multicollinearity can not be eliminated. This paper indicates that the multicollinearity in independent variable sets can evidently affects the deriving of components and the regression parameters of PLS1 regression model. This can result in that the dependent variable can not be properly explained by components. Meanwhile, the regression model and the components may be difficult to understand. Therefore, it is still necessary to select independent variables carefully before building PLS1 models.

#### Acknowledgment

The authors acknowledge the financial supports from National Natural Science Foundation of China under the grant number 70531010, number 70521001, and number 70371007, and the financial supports from Beijing Natural Science Foundation under the grant number 9052006.

#### References

- [1] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method[C]. Ruhe A, Kågström B (Eds), Proc. Conf. Matrix Pencils, Lectures Notes in Mathematics. Heidelberg: Springer-Verlag, 1983.
- [2] Tenenhaus M. L'approche PLS[M]. Revue de Statistique Appliquée, 1999.
- [3] Rolf Ergon, Reduced PCR/PLSR models by subspace projections[J]. Chemometrics and Intelligent Laboratory Systems, 2006-03 Volume: 81: 68-73
- [4] Bjørn-Helge Mevik; Henrik René Cederkvist, Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR) [J]. Journal of Chemometrics, 2004-09 Volume: 18 Page: 422-429
- [5] Rolf Ergon, Constrained numerical optimization of PCR/PLSR predictors [J]. Chemometrics and Intelligent Laboratory Systems, 2003-02 Volume: 65 Page: 293-303
- [6] Martin Petersen, Marianne Dyrby, Søren Toubro, Søren Balling Engelsen, Lars Nørgaard, Henrik Toft Pedersen, and Jørn Dyerberg, Quantification of Lipoprotein Subclasses by Proton Nuclear Magnetic Resonance-Based Partial Least-Squares Regression Models[J]. Clin. Chem. 2005 51: 1457-1461.