Journal of Data Science **4**(2006), 257-274

An Modified PLSR Method in Prediction

Bo Cheng[1] and Xizhi Wu[2]
[1]*Altair Engineering Company and* [2]*Renmin University of China*

Abstract: Among many statistical methods for linear models with the multicollinearity problem, partial least squares regression (PLSR) has become, in recent years, increasingly popular and, very often, the best choice. However, while dealing with the predicting problem from automobile market, we noticed that the results from PLSR appear unstable though it is still the best among some standard statistical methods. This unstable feature is likely due to the impact of the information contained in explanatory variables that is irrelevant to the response variable. Based on the algorithm of PLSR, this paper introduces a new method, modified partial least squares regression (MPLSR), to emphasize the impact of the relevant information of explanatory variables on the response variable. With the MPLSR method, satisfactory predicting results are obtained in the above practical problem. The performance of MPLSR, PLSR and some standard statistical methods are compared by a set of Monte Carlo experiments. This paper shows that the MPLSR is the most stable and accurate method, especially when the ratio of the number of observation and the number of explanatory variables is low.

Key words: Modified partial least squares regression (MPLSR), multicollinearity, partial least square regression (PLSR), ridge regression (RR), principal components regression (PCR), variable subset selection method (VSS).

1. Introduction

In automobile market, the auction price of two-year-in-service vehicle is an important indicator of that vehicle's market value, which is of great interest to manufacturers, dealers, financial institutions and consumers. When linear model is used to predict the auction price, multicollinearity arises. Multicollinearity often exists when the number of explanatory variables is large compared to the number of observations, and it causes difficulty estimating parameters.

To solve multicollinearity problem, many statistical methods have been suggested. The variable subset selection method (VSS) is used to avoid the multicollinearity caused by too many variables, and the stepwise version is used here. The ridge regression (RR) was suggested by Hoerl and Kennard (1970) as a method for stabilizing regression estimates in the presence of multicollinearity, which assumes that the regression coefficients are not likely to be very large. Principal components regression (PCR), introduced by Massy (1965), tries to reduce the dimension and avoid multicollinearity by using just a few components, the linear combinations of the explanatory variables.

Being a comparatively new method, the partial least squares regression (PLSR) has became the most popular regression method in chemometrics. PLSR was suggested by Wold (1975), Wold *et al.* (1984), Martens (1985, 1989), Helland (1988) and Garthwaite (1994). The PLSR can be traced from general systems-analysis methods of Wold. It is a useful tool when multicollinearity exists among explanatory variables and when the number of explanatory variables is very large compared to the number of observations. PLSR has been studied in great details. Frank (1993) and Goutis (1996) proved properties of PLSR estimates. Ruscio (2000) studied the relationship between the PLSR algorithm for univariate data and Cayley-Hamilton polynomial expression. Stone (1990) introduced continuum regression based on OLS, PLSR and PCR. Goutis (1996) introduced a modification of PLSR using a roughness penalty. Wold (1992) and Durand (1997) extended PLSR into nonlinearity using spline functions. Presently PLSR have been applied in many fields, especially broadly in chemistry as “the use of mathematics and statistics on chemical data” (Martens, 1989). PLSR have been compared with other methods in chemistry, see Phatak (1993), Phatak (1997) and Ter Braak (1998). PLSR was also combined with neural network as a new subject in nonlinear analysis (Ham, 1997). The software for the PLSR regression is available in some packages such as Unscrambler 7.5 (a PLSR and experimental design software), SAS and SIMCA 8.0 (a PLSR software).

When the four methods (PLSR, RR, PCR and VSS) are used to predict the auction price referred in the first paragraph, although the algorithms of PLSR and RR reach better results compared to the very large average relative errors from using VSS and PCR, their performances on five different vehicle lines are unstable, and therefore unsatisfactory, despite the fact that the five vehicle lines have very similar position in automobile market.

While studying this practical question, we discovered the reason behind the unstable performance of PLSR and developed the more stable modified partial lease square regression (MPLSR), a modification of PLSR.

This paper is organized as follows. Section 2 provides the background of the practical problem, predicting auction price, and the results from using the four methods (PLSR, RR, PCR and VSS). Section 3 presents the idea of the PLSR method and analyzes its shortcoming, which motivates us to introduce a new MPLSR method. Section 4 introduces the algorithm of the new MPLSR method. Section 5 applies the MPLSR in our practical problem on automobile market and compares its prediction results with those of other four methods. Section 6 presents a simulation study to compare the performance of MPLSR, PLSR, VSS, RR and PCR.

2. Automobile Market Prediction Results

In the automobile market, the auction price of a two-year-in-service (2YIS) vehicle is of special interest because it is the base of many important decisions. For example, it is used to calculate the lease end value. When a consumer leases a vehicel on January 2005 for 2 years, he will return the vehicle on January 2007. The manufacturer suggested price minus the lease end value is his payment for 2 years lease. In this case, manufactor needs to know the auction price of a 2YIS vehicle on January 2007. The auction price of 2YIS vehicles is highly correlated with the quality of the vehicle. A vehicle with good style and durabality will be fetched a good price. On the other hand, if a vehicle is a trouble maker, it has less chance to be sold at a good price. The Compact Utility segment is one of the most popular segments in the United States. It has attracted a lot of attention recently. We select five major vehicles from this segment: Explorer, 4Runner, Grand Cherokee, Cherokee and Blazer. Our goal is to predict their auction price of 2YIS vehicles.

The data used in the study includes the auction prices and twenty major factors (indexes) including APEAL Score (APEAL is referred to as Automotive Performance, Execution and Layout) measuring an owners* delight with the design and features of their vehicle, customer satisfactory indexes, durability indexes, money against market (incentive), manufacturer suggest price (MSRP), style age and used-car consumer price index (UCPI). On the auto market, the manufacturers modify their vehicles and introduce new model year vehicles each year. For example, on October 1998 the manufacturers introduced 1999 model year vehicles that are modified based 1998 model year vehicles. Most modifications are minor. But some modifications are major that are called major refreshing. In major refreshing the exterior styling and interior styling are changed. The style age equals current model year subtract the model year of last major refreshing.

The OLS method is used at first and only three independent variables are significant under the t-test with the coefficient of determination R-square greater than 0.8 while the all twenty independent variables are present. If only the three significant variables are used as independent variables, the R-square is only 0.3. Collinearity is naturally suspected, and among the twenty condition indexes, nine are larger than 60 and one is greater than 1000. Collinearity can be also expected by just looking the original meanings of these independent variables. For example, the vehicle lines with higher APEAL and better durability will have higher MSRP, lower incentive and higher customer satisfactory index.

Our study particularly covers two-year-in-service leased vehicles. The auction prices are the auction prices from automakers to dealers. Linear model is built on this data set. The auction price of one kind of vehicle line from January 1995 to June 1999 is the response variable, and all twenty other major variables of this kind of vehicle line from the corresponding two-year-in-service periods, which is from January 1993 to June 1997, are explanatory variables. Here, the auction price on January 1995 and the values of other variables on January 1993 are from the same batch of vehicles because the auction price of a new vehicle produced on January 1993 become available only after two years. This linear model is for capturing the relationship between vehicle's attributes and its auction price two years later.

The Regression ARIMA is the first model we try to use in this study. But this is a long term (24 months) forecast and the multicollinearity makes the problem complicated, the time series method doesn't have an advantage.

The methods of VSS(stepwise), RR, PCR and PLSR are natural candidates. Firstly all the four methods are applied to the five kinds of compact utility vehicles, and the prediction results are analyzed.

The monthly average of auction prices from July 1999 to December 2000 (18 months), not used in regression, are used to verify the prediction result. For analyzing the prediction results, we calculate the errors between predicted and the actual auction price. The average of the relative prediction errors (ARE) $\sum_{t=1}^{18} |y_t - \hat{y}_t|/(18\bar{y})$ is used as a criterion to test the predicting capability of a model. Here, $\bar{y}$ is the mean of the auction prices from January 93 to June 99, y_t is the actual auction price and $\hat{y}_t$ is the predicted auction price. The relative errors $(y_t - \hat{y}_t)/\bar{y}$, $t = 1, ..., 18$ from using the four methods are plotted together for comparison. Because the ARE is a commonly used index in practical, it is used here, while we will use a similar measure, statistic average prediction squared error (PSE), in Section 6. Table 1 shows the average relative prediction errors of the five kinds of SUV vehicle lines using all the four methods. The results show that the PLSR is the best method among five methods.

Table 1: Average relative errors of the four methods

Methods	Vehicles				
	Explorer	4Runner	Cherokee	Grand-Cherokee	Blazer
PLSR	4.4%	2.4%	3.3%	6.4%	14.6%
VSS(Stepwise)	11.4%	9.7%	16.7%	26.7%	21.8%
PCR	9.4%	7.9%	26.9%	17.1%	16.5%
RR	4.5%	3.9%	5.6%	17.4%	8.9%
(ridge parameter)	(0.008)	(0.34)	(1)	(1)	(0.14)

Here, the one-at-a-time cross validation is used to select the cut-off place of the PCR. The independent matrix is standardized in RR. For the detail information about PLSR, please see the Appendix. Since one of the important assumptions of RR is that the regression coefficients are not likely to be very large. So the ridge parameter is usually selected from the range [0,1] in practical problems. In Table 1, the ridge parameters are the optimal ones in [0,1], which provide the lowest ARE. From Table 1, the other three methods produce larger predicting errors than the PLSR on average.

However, the performance of PLSR is inconsistent among the five kinds of vehicle lines. It obtains satisfactory prediction result for first three kinds of vehicle lines but not the last one. When PLSR method is used, the prediction auction prices of Blazer have large bias from its actual value. It is this case that causes our attention. Discovering the reason that the PLSR becomes inefficient in this case may lead to the key of overcoming the shortcoming of PLSR method. Figure 1 shows the relative errors in predicting auction prices of the five kinds of SUV vehicle lines.

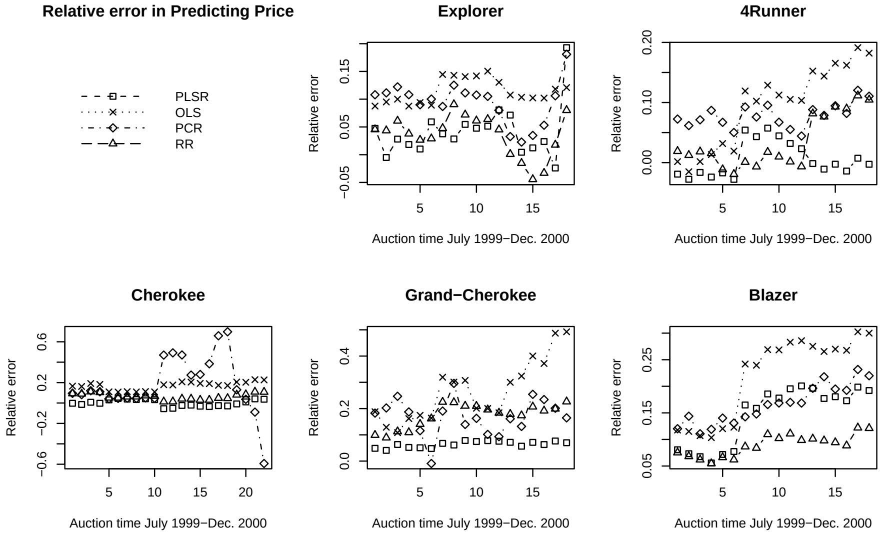

Figure 1: Relative errors in predicting of auction price

From Figure 1, the predicted values of Blazer from all the four methods are much higher than the actual auction value. This inefficiency of all methods may be caused by the irrelevant-to-the-response information contained in explanatory variables during the prediction period.

3. The Situation Where the PLSR Method Does Not Work Well

As we have seen in the last section, PLSR method does not work well in all situations. It provides a very inaccurate prediction of auction price for Blazer although the prediction results of other 4 vehicle lines are reasonable. This phenomenon is caused by the irrelevant information in the explanatory matrix to the response variable.

Let the linear model (here only univariate response is considered) be

$$Y = X\beta + \epsilon,$$

where Y is an $n \times 1$ response vector and X a known $n \times k$ explanatory matrix, and ϵ is a noise term with the same dimensions as Y. Matrix X of explanatory variables contains two types of information. One type is relevant to the response variable Y and therefore useful in predicting the value of Y. The other type is irrelevant to Y and hence causes inefficiency in the prediction. The idea of PLSR algorithm is to extract components (factors) $\{t_i\}$ from X, which are relevant to Y. These components are extracted in decreasing order of relevance measured by covariance $\text{Cov}(t_i, Y)$. Let T be the matrix of the selected components t_i's, and therefore $T = XW$, where the columns of W are weight vectors for the X columns. Then ordinary least squares procedures for the regression of Y on the matrix T are performed to produce the coefficient vector V or $\hat{Y}_{PLS} = TV$. Then the estimator $\hat{\beta}_{PLS}$ of the original β has the form of $\hat{\beta}_{PLS} = WV$. A version of detailed PLSR algorithm is provided in Appendix.

In PLSR, despite different approaches, each factor t_i is selected to maximize, in the sense of absolute values, the covariance $\text{Cov}(t_i, Y)$, where

$$\text{Cov}(t_i, Y) = \text{Corr}(t_i, Y)\sqrt{\text{Var}(t_i)\text{Var}(Y)} \propto \text{Corr}(t_i, Y)\sqrt{\text{Var}(t_i)}.$$

Since the t_i is a linear combination of independent variable, its variance may not be 1; the variances of all independent variables are standardized as 1. An ideal situation is that *both* $\text{Var}(t_i, Y)$ and $\text{Corr}(t_i, Y)$ decrease monotonously as $\text{Cov}(t_i, Y)$ decreases during the process of selecting components; that means the most representative (due to large variance) and the most relevant (due to large correlation) elements in X would be used for the regression. Unfortunately, it is not always true that a large $\text{Cov}(t_i, Y)$ will guarantee that *both* $\text{Corr}(t_i, Y)$ and $\text{Var}(t_i)$ are all large at the same time. It is possible that a factor t_i corresponding to a large $\text{Cov}(t_i, Y)$ caused by a large $\text{Var}(t_i)$ but relatively smaller $\text{Corr}(t_i, Y)$ may be selected while another factor t_i^* with a slightly smaller $\text{Cov}(t_i^*, Y)$ caused by a relatively smaller $\text{Var}(t_i^*)$ but a larger $\text{Corr}(t_i^*, Y)$ may be discarded.

As a consequence of discarding information relevant to Y, $\hat{Y}_{PLS}$ has lower correlation with Y. Let us see a simple illustrative example, where no error term is added for the obviousness.

Example 1. Suppose our data is

$$Y = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 20 & 1 & 0 & 10 \\ 65 & 0 & 1 & 9 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

or $X = (z_1 + 20z_2 + 65z_3,\ z_2,\ z_3,\ 10z_2 + 9z_3)$ where z_i denotes 4×1 vector which the ith element is one and the others are zero. We want to make regression of Y on X. The solution of the regression is obvious: in the term of relation between X and Y, $Y = 3z_1$. Obviously the ordinary least squares (OLS) method does not work due to the multicollinearity. What would PLS method say on this example? We use PLS to find the factors t_1, t_2 and t_3 according to their values of $\mathrm{Cov}(t_i, Y)$ in descending order. The values of $\mathrm{Cov}(t_i, Y)$, $\mathrm{Var}(t_i)$ and $\mathrm{Corr}(t_i, Y)$ for $i = 1, 2, 3$ are

	$\mathrm{Cov}(t_i, Y)$	$\mathrm{Var}(t_i)$	$\mathrm{Corr}(t_i, Y)$
t_1	1.505	3.154	0.565
t_2	0.221	2.102	0.102
t_3	0.015	0.00014	0.819

With common criterion in cross-validation, the PLS method selects only t_1 to be the regressor because it has the largest variance $\mathrm{Cov}(t_1, Y)$, which however is almost entirely due to the largest $\mathrm{Var}(t_1)$ despite its small $\mathrm{Corr}(t_1, Y)$. The reason for having these values of covariance, variance and correlation is the composition of t_i. With matrix notation, the relation between t_i and z_j are

$$\begin{pmatrix} t_1 & t_2 & t_3 \end{pmatrix} = \begin{pmatrix} \mathbf{1}, z_1 & z_2 & z_3 \end{pmatrix} \cdot \begin{pmatrix} 1.525 & 0.208 & -0.015 \\ -0.020 & 0.013 & 0.029 \\ -2.669 & -2.172 & 0.015 \\ -3.412 & 1.327 & 0.014 \end{pmatrix}.$$

Clearly the chosen t_1, which mainly composed with z_2 and z_3 through the last three columns of Z, has little relation with Y or z_1. On the contrary, the last factor t_3 which has no chance to be selected by Cross-validation even under the least conservative criterion because of its small covariance although it is more correlated with Y than the first two. Therefore the PLSR does not work in this situation. To emphasize the information relevant to Y in the modeling process in order to reach better prediction, next we introduce the following modified partial least squares regression (MPLSR) algorithm.

4. Modified Partial Least Squares Regression (MPLSR) Algorithm

The main idea of our MPLSR methods is to use an orthogonal projection that removes from $\hat{Y}_{PLS}$ the elements irrelevant to Y. First, we find some factors which are linear combination of independent variables and orthogonal with Y. Second, the effect of irrelevant information in X are removed by projecting the $\hat{Y}_{PLS}$ on orthogonal complement space of those factors. The following is the algebra of the MPLSR algorithm.

For our model $Y = X\beta + \epsilon$, since $X^{'}YY^{'}X$ is a real symmetric matrix with rank 1, it has $k-1$ orthogonal eigenvectors correspondent to the zero eigenvalue. Let $b_1, ..., b_{k-1}$ denote the $k-1$ eigenvectors corresponding to zero eigenvalue and $B \equiv (b_1, ..., b_{k-1})$, a $k \times (k-1)$ matrix with columns of $b_1, ..., b_{k-1}$. Because $b_i^{'}X^{'}YY^{'}Xb_i = 0$, or $Y^{'}Xb_i = 0$ for $i = 1, ..., k-1$, the k-vectors $\{b_1, ..., b_{k-1}, X^{'}Y\}$ form an orthogonal basis of a k-dimensional space. All those orthogonal to Y can be expressed as $XB\alpha$. Among unit vectors α ($\alpha^{'}\alpha = 1$), we pick up those that make variance of $XB\alpha$ maximum, which are the eigenvectors corresponding to the maximal eigenvalues of $B^{'}X^{'}XB$. So we select a number of maximal eigenvalues of $B^{'}X^{'}XB$, $\lambda_1, ..., \lambda_s$, such that the cumulative eigenvalue contribution proportion $\sum_{i=1}^{s} \lambda_i / \sum_{i=1}^{k-1} \lambda_i$ is greater than a certain value, 99% say. Let their corresponding eigenvectors be columns of matrix $A \equiv (\alpha_1, ..., \alpha_s)$. Also let $U = XBA$, which is orthogonal to Y. The projection of X orthogonal to U is $(I_n - P_U)X \equiv X - U(U^{'}U)^{-1}U^{'}X = X(I_k - BA(U^{'}U)^{-1}U^{'}X) = XD$ with $D \equiv (I_k - BA(U^{'}U)^{-1}U^{'}X)$. Let the original PLSR fitted vector be $\hat{Y}_{PLS}$ and the estimated coefficient vector be $\hat{\beta}_{PLS}$. Then the fitted value from our MPLSR algorithm is defined by $\hat{Y}_{MPLS} \equiv XD\hat{\beta}_{PLS}$. Let the estimated coefficients by MPLSR be $\hat{\beta}_{MPLS} = D\hat{\beta}_{PLS}$.

The estimation $\hat{Y}_{MPLS}$ reduces from $\hat{Y}_{PLS}$ the element of irrelevant information to Y and emphasizes the roles of relevant information during the estimation process.

Since the MPLSR is based on the result of PLSR method, a better result will be obtained when the prediction of Y by using PLSR methods includes more relevant information.

5. A Comparison of the MPLSR with Four Methods for Auction Price Case

In this section, we continue the discussion of the prediction problem in Section 2. The proposed MPLSR is used to predict residual values of two-year-in-service vehicles, and the results are compared to those from using the PLSR, VSS, RR and PCR methods. Table 2 presents average relative errors of predicting the five kinds of SUV vehicle lines using all the five methods (the results except MPLSR have been shown in Table 1).

Table 2: Average relative errors of the five methods

Methods	Vehicles				
	Explorer	4Runner	Cherokee	Grand-Cherokee	Blazer
MPLSR	**1.64%**	**3.05%**	**3.6%**	**5.1%**	**1.7%**
PLSR	4.4%	2.4%	3.3%	6.4%	14.6%
VSS (stepwise)	11.4%	9.7%	16.7%	26.7%	21.8%
PCR	9.4%	7.9%	26.9%	17.1%	16.5%
RR	4.5%	3.9%	5.6%	17.4%	8.9%
(ridge parameter)	(0.008)	(0.34)	(1)	(1)	(0.14)

From Table 2, the other four methods produce larger predicting errors than the MPLSR on average. Comparing to PLSR, the MPLSR method produces smaller error, except that for 4Runner and Cherokee, which are very close, and both methods produce similar error patterns that closely track each other along different time (see Figure 2).

Unlike the PLSR, the performance of MPLSR is consistent when it is used to predict the auction prices of the five similar vehicle lines. When the MPLSR method is used, the average relative errors of the five vehicles are almost under 5%, and MPLSR's result for Blazer is much better than those from the other four methods.

The MPLSR method has a consistent performance not only on predicting auction prices of different kinds of vehicle lines but also on predicting one kind of vehicle's auction prices on different time. Table 3 provides the standard deviation of predicting errors that measure the deviation level of predicting errors of one kind of vehicle line.

Table 3: Standard deviation of predicting errors of the five methods

Methods	Vehicles				
	Explorer	4Runner	Cherokee	Grand-Cherokee	Blazer
MPLSR	0.016	0.036	0.026	0.011	0.017
PLSR	0.047	0.029	0.036	0.012	0.056
VSS(stepwise)	0.02	0.07	0.045	0.117	0.08
PCR	0.04	0.02	0.31	0.07	0.04
RR	0.04	0.02	0.028	0.045	0.02

As noted in Table 3, in most situations, the standard deviation of errors from using MPLSR is less than that from using the other four methods on average for the test data. That demonstrates the better predicting capability and more stable results of MPLSR compared to the other methods.

Figure 2 presents the relative errors in predicting auction prices of the five SUV vehicle lines. In each picture, the relative errors from using the five methods are put together for comparison. As shown in Figure 2, the errors of predicting auction prices of Explorer by using the PLSR method are particularly large in the last two months. This unexpected large error is partially caused by the noise presented in the original data. The error is, however, significantly smaller when the MPLSR method is applied. The MPLSR removes irrelevant information and reduces the disturbance caused by noise. The errors of predicting auction price using the MPLSR method have smaller fluctuation.

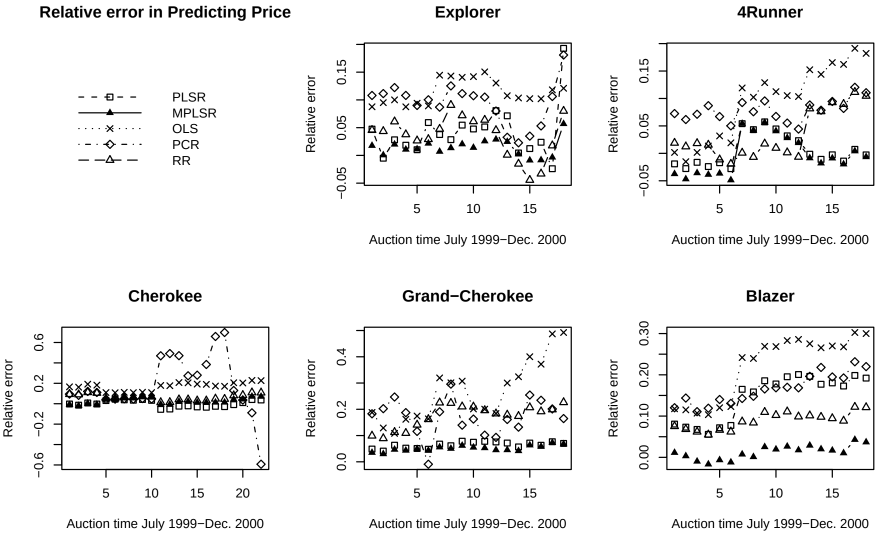

Figure 2: Relative errors in predicting of auction price (with MPLSR)

The predicted results of 4Runner's auction price, using PLSR and MPLSR, are very close in the year 2000. The average of the errors by using PLSR is smaller than that of MPLSR, and both methods result in similar patterns.

It is clear that the MPLSR produces significantly better-predicted results for Blazer than the other methods. The predicted values of PLSR, VSS, PCR and RR are much higher than the real auction value. All the four methods are influenced by the same kind of irrelevant information, and this information becomes very abnormal than usual in the last half year. Without removing the irrelevant information, the PLSR produces results having a large bias, and due to the removal of irrelevant information, the MPLSR's results track the trends very well although the pattern of prediction errors is similar to that of PLSR.

For each of these five vehicles, the errors by the PLSR and the MPLSR follow the same trend in time series, although their magnitudes appear to be different. Because the MPLSR emphasizes the information relevant to Y, its predicted results often follow the real values more closely than those of the PLSR method.

The five methods are also used to predict auction prices of five upper middle vehicles. The results are similar. The MPLSR method provides the most accurate and stable predicting auction prices among the five methods.

This practical example demonstrates that MPLSR algorithm does have advantages over other four when the multicollinearity exists. For further investigation, next we use Monte Carlo analysis to compare MPLSR method with the others.

6. A Simulation Comparison of MPLSR, PLSR, VSS, RR and PCR

To understand in what situations MPLSR can be expected to work well compared to other standard methods, VSS (stepwise), RR, PLSR and PCR, a set of Monte Carlo experiments is performed, and a summary of the results is presented in this section.

The five methods are compared in 360 different situations with different numbers of explanatory variables (k = 30, 60 and 100) and different levels of collinearity in the explanatory matrix. This means that the correlation matrix of explanatory variables have different structures (low collinear-all off-diagonal elements 0.4; middle collinear-all off-diagonal elements 0.7; high collinear-all off-diagonal elements 0.9). These situations also have different noise-to-signal ratio $\{\sigma/\mathrm{Var}(\alpha' X)^{1/2} = 0.05 \text{ or } 0.1\}$ and different true regression coefficients (20 sets of different regression coefficients β are generated randomly from normal distribution $N(0, 100)$). So there are totally $3 \times 3 \times 2 \times 2 = 360$ situations studied. For each situation, 100 data sets are generated and the results are reported as means of the 100 replications. Each data set includes 150 observations. The first 50 observations are training data that is used to estimate the regression coefficients by using the five regression methods (MPLSR, PLSR, VSS, RR and PCR) respectively. The last 100 observations are test data that is used to test the performance of the five methods.

The average prediction squared error (PSE) over the 100 test observations is used as the statistic to compare the performance of the five regression methods:

$$PSE = \frac{1}{100} \sum_{i=51}^{150} (y_i - \hat{y}_i)^2.$$

The PSE values for each method in each situation are averaged over the 100 replications to compare the predicting capabilities of the five methods in different situations. The SDP values (standard deviation of PSE) for each method in each situation are calculated over the 100 replications to measure the stability of prediction of the five methods. The lower values of PSE and SDP indicate the better performance of the corresponding method.

Frank & Friedman (1993) provided the optimal ridge parameter λ that minimizes the mean squared error (MSE) of the prediction, and therefore the optimal ridge parameter can be calculated in each situation to be the base of RR. The results of the four regression methods (PLS, VSS, RR and PCR) are obtained by SAS standard procedure (PROC PLS and PROC REG).

The results of the Monte Carlo experiments are presented in Figures 3-6. Figure 3 shows the average PSE and SDP of the five methods over the 360 situations.

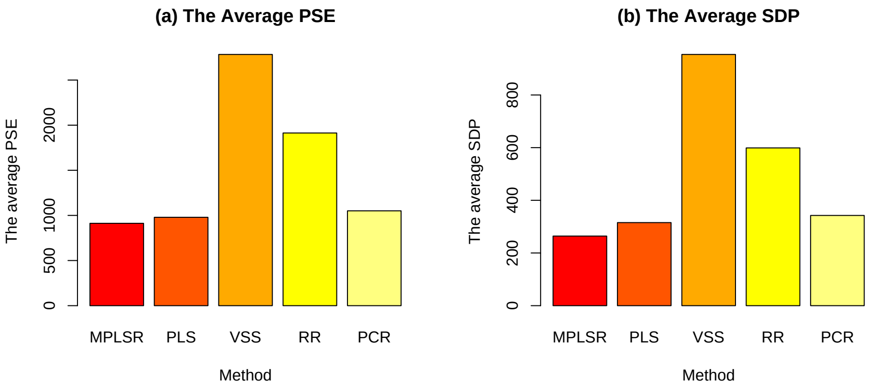

Figure 3: Average PSE (3.a) and SDP (3.b) of five methods over 360 situations.

Figure 3 demonstrates that our new method MPLS has the best performance with smallest average PSE and SDP, which means accuracy and stability, and VSS (stepwise) being the worst. Table 4 provides the percentages that the MPLS method reduces the values of PSE and SDP from four other methods. Here in the table PPSE and PSDP are

$$PPSE = \frac{PSE \text{ of compared method} - PSE \text{ of } MPLSR}{PSE \text{ of compared method}},$$

$$PSDP = \frac{SDP \text{ of compared method} - SDP \text{ of } MPLSR}{SDP \text{ of compared method}}.$$

Table 4: Reduced percentages based on all situations

Percentage	Method			
	PLS	VSS	RR	PCR
PPSE	6.8%	67%	52%	13%
PSDP	16%	72%	56%	23%

From Table 4, one can see that having the the smallest PSE among the five methods, MPLSR reduces PSE by 6.8 percents compared to PSE of PLSR, which has the second smallest PSE. This means that the MPLSR improves the predicting capability significantly. Since the SDP of MPLSR is the lowest among the five methods, MPLSR is the most stable method in the five methods. When the MPLSR is used, the SDP improves by 16 percents compared to the result of using PLSR method. The advantage of MPLSR on stability of prediction is significant compared with the other four methods.

From Figure 3 and Table 4, the new method MPLSR provides the best average overall performance significantly; the PLSR and PCR follow closely, and RR gives an inferior overall performance just slightly better than VSS. Since the performances of these methods may change with different situations, discussion their performance in different situation is necessary.

Figures 4-6 present a graphical detailed summary of classified results from this simulation analysis according to the characteristics of three kinds of data characteristics (number of independent variables, collinear level and noise-to-signal ratio).

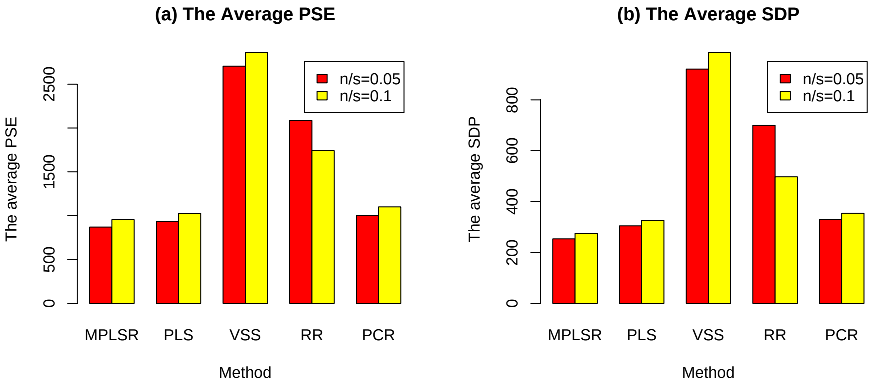

Figure 4: Performance comparison of five methods on PSE (4-a) and SDP (4-b) for two levels of noise-to-signal ratio.

Figure 4 demonstrates MPLSR provides the best results in these levels of noise-to-signal ratio. Also RR behaves in a different way from the others: when noises increase, RR's predicting capability and stability level increases while those of MPLSR, PLSR, VSS and PCR decrease.

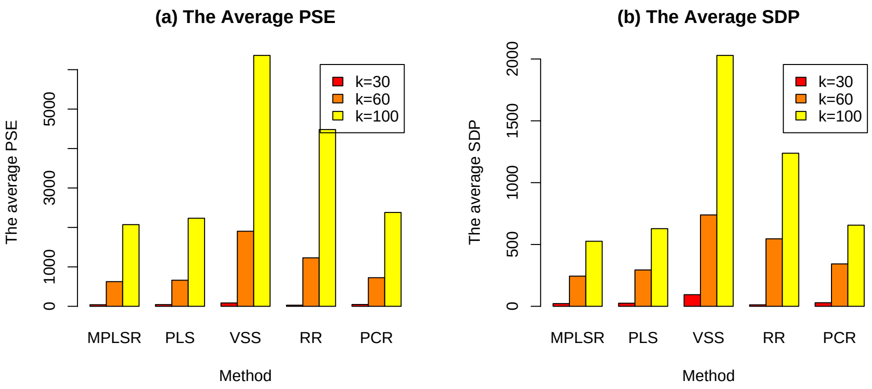

Figure 5: Performance comparison of five methods on PSE (5-a) and SDP (5-b) on $k = 30, 60$ and 90.

Figure 5 shows that the result of MPLSR is the best except when $k = 30$. In the situation where $k = 30$, RR (Note: the optimal parameter λ of RR is known in the simulation while λ is almost impossible to know in a real problem) performs slightly better than MPLSR. But this advantage of RR disappears rapidly when the ratio of the number of observations and the number of explanatory variables (OVR) decreasing. the values of PSE increase sharply when the number of explanatory variables increases.

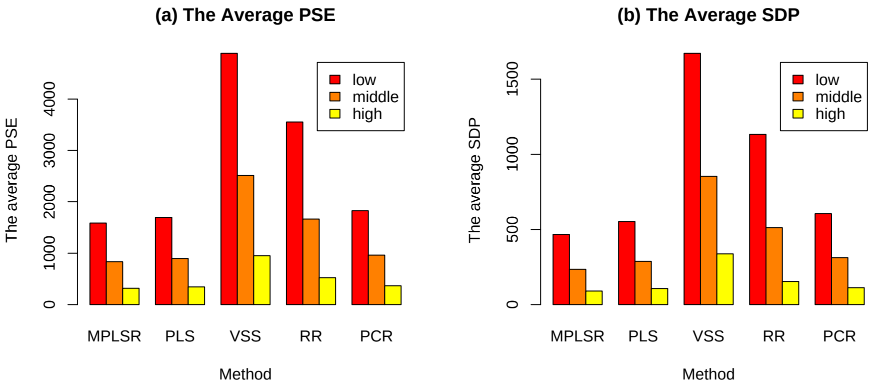

Figure 6: Performance comparison of five methods on PSE (6-a) and SDP (6-b) on low, middle and high multicollinear situations.

Figure 6 shows that for the three kinds of collinearity levels, MPLSR gives the best predicting model. When the collinearity increases, the advantage of MPLSR method becomes more prominent.

Figure 4-6 shows that MPLSR provides the best and the most stable predicting results (the lowest PSE and SDP) among the five methods in almost all situations except one situation (see Figure 3, when $k = 30$), where the performance of the RR is better.

However, in our automobile market example, the predicting results of RR are no better than MPLSR although the OVR is higher than 50/30. One of the reasons is the difficulty of determining the ridge parameter in practice because it is impossible to obtain the optimal ridge parameter in a real problem. Because the RR method is sensitive to the ridge parameter, a bad ridge parameter will produce a model that cannot obtain a reasonable prediction. From this point of view, MPLSR is a more practicable method than the RR.

We should notice that the pattern of the five methods are similar in both the practical example and in simulations. This ensures the advantage of MPLSR method in different situations.

7. Discussion

In this paper, MPLSR method has been introduced when the explanatory matrix X includes much information irrelevant to the response variable Y. It is an algebraic algorithm based on the result of the PLSR method. Both Monte Carlo experiments and the practical example demonstrate that the new method produces more accurate and stable results than other standard statistical methods (VSS, RR, PCR and PLSR), especially when the observations-variables number ratio is low and the multicollinearity is high among independent variables.

We suggest that even in the steps of selecting components in PLSR, one should select not only the components with large covariance with the dependent variable Y but also the components with large correlation with variable Y.. One possible way is to use PLSR between Y and XD instead of between Y and X; another is to make compromise between $\mathrm{Var}(t_i, Y)$ and $\mathrm{Corr}(t_i, Y)$ in the criterion used for selecting components in PLSR process.

Acknowledgements

This work is supported by National Science Foundation of China(10431010), Key Project of Key base of Wducation Ministry of China (05JJD910001) and Center for Applied Statistics of Renmin University.

Appendix: The Algorithm of PLSR Method

Let Y be an $n \times 1$ dependent vector and X a known $n \times p$ explanatory matrix. Assume X and Y are standardized. The linear model is $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$.

The following is a brief description of the PLSR algorithm (see Helland, 1988). Let $P_s = S(S^{'}S)^{-1}S^{'}$ denote the projection matrix onto the space spanned by column vector(s) of a matrix S.

(1) Selection of orthogonal component $\{t_i\}$

(a) Calculating $\{t_i\}$ sequentially

Let $X_0 = X$ initially. The component t_k is selected in step k and calculated as:

$$w_k = X_{k-1}^{'}Y/\sqrt{Y^{'}X_{k-1}X_{k-1}^{'}Y},\ t_k = X_{k-1}w_k \ \text{ and } \ p_k = X_{k-1}^{'}t_k/t_k^{'}t_k \qquad (1)$$

and $X_k = X_{k-1} - t_k p_k^{'} = (I - P_{t_k})X_{k-1}$.

(b) Using cross-validation (CV) criterion at step k

Here, the CV (Stone, 1974) is used as the criterion to decide whether the component k should be selected into the model. Let T^k be the matrix which has k columes and the i-th colume is t_i. Let SSE^k be the residual sum of the regression model where the dependent variable is Y and explanatory matrix is T^k, and the partial SSE, or PSSE, is determined by

$$PSSE^{k+1} = \sum_{i=1}^{n}(y_i - T_{\{i\}}^{k+1}(T_{(i)}^{k+1\,'}T_{(i)}^{k+1})^{-1}T_{(i)}^{k+1\,'}Y_{(i)})^2 \qquad (2)$$

where symbol $T_{\{i\}}^{k+1}$ denotes the i-th row of T^{k+1}, $T_{(i)}^{k+1}$ represents T^{k+1} without its i-th row, and $Y_{(i)}$ is Y without its i-th element y_i.

At step k the CV statistic Q^k is defined as $Q^k = 1 - PSSE^{k+1}/SSE^k$. If $Q^k \geq 0.0975$, proceed with step $k+1$; otherwise the selection process stops.

(2) Regression on chosen T^q

With selected q components $T^q = (t_1, ..., t_q)$, the regression model where Y is the dependent variable and T^q is the explanatory matrix is fitted by OLS model

$$Y = T^q r^q + \epsilon = r_1 t_1 + \cdots + r_q t_q + \epsilon \qquad (3)$$

(3) and the estimation of the coefficient r^q is obtained as $\hat{r}^q = (T^{q'}T^q)^{-1}T^{q'}Y$. Then the estimation of Y is $\hat{Y} = \hat{r}_1 t_1 + \cdots + \hat{r}_q t_q$. From (1),

$$t_i = X_{i-1}w_i = X\sum_{j=1}^{i}(I_p - w_j p_j^{'}).$$

Let $w_i^* = \sum_{j=1}^{i}(I_p - w_j p_j^{'})$ and substitute t_i in (3) with Xw_i^*, estimation of Y in PLS is

$$\hat{Y}_{PLS} = \hat{r}_1 t_1 + \cdots \hat{r}_q t_q = \hat{r}_1 X w_1^* + \cdots \hat{r}_q X w_q^* = X \sum_{i=1}^{q} \hat{r}_i w_i^*. \tag{4}$$

For simplicity, let $\hat{\beta}_{PLS} = \sum_{i=1}^{q} \hat{r}_i w_i^*$, then $\hat{Y}_{PLS} = X\hat{\beta}_{PLS}$.

References

Durand, J., and Sabatier, R., (1997). Additive splines for partial least squares regression. *Journal of American Statistical Association* **92**, 1546-1554.

Frank, L. E., and Friedman, J. H., (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

Garthwaite, P. H., (1994). An interpretation of partial least squares. *Journal of American Statistical Association* **89**, 122-127.

Goutis, C., (1996). Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics* **24**, 816-824.

Goutis, C., and Fearn, T., (1996). Partial least squares regression on smooth factors. *Journal of American Statistical Association* **91**, 627-632.

Helland, I. S., (1988). On the structure of partial least squares regression. *Commun. Statist. -Simula.* **17**, 581-607.

Hoerl, A. E., and Kennard, R. W., (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **8**, 27-51.

Martens, H., (1985). *Multivariate Calibration.* Dr. techn. Thesis. Technical University of Norway, Trondheim.

Martens, H., and Næs, T., (1989). *Multivariate Calibration.* Wiley.

Massy, W. F., (1965). Principal components regression in exploratory statistical research. *Journal of the American statistical Association* **60**, 234-246.

Phatak, A., (1993). *Evaluation of Some Multivariate Methods and Their Applications in Chemical Engineering.* Ph.D. thesis, University of Waterloo.

Phatak, A., and de Jong, S., (1997). The geometry of partial least square. *Journal of Chemometrics* **11**, 311-338.

Ruscio, D. D., (2000). A weighted view on the partial least-squares algorithm. *Automatica* **36**, 831-850.

Stone, M., (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Ser. B* **36**, 111-147.

Stone, M., and Brooks, R. J., (1990). Continuum regression: Corss-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *Journal of the Royal Statistical Society, Ser. B* **52**, 237-269.

Ter Braak, C. J., and de Jong, S., (1998). The objective function of partial least squares. *Journal of Chemometrics* **12**, 41-54.

Wold, H., (1975). Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett* (Edited by J. Gani). Academic Press.

Wold, H., and J reskog, K. G., eds., (1982). *Systems under indirect observation. Causality - structure - prediction.* Contributions to Economic Analysis **139**, parts I and II. North-Holland.309

Wold, S., Wold, H., Dunn, W. J. and Rune, A., (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Siam J. Sci. Stat. Comput.* **5**, 735-743.

Wold, S., (1992). Nonlinear partial least square modeling. II. spline inner relation. *Chemometrics and Inteligent Laboratory Systems* **14**, 71-84.

Received February 5, 2005; accepted March 11, 2005.

Bo Cheng
Altair Engineering, Inc.
1820 E. Big beaver
Troy, MI 48087
USA

Xizhi Wu
School of Statistics
Renmin University of China
Beijing 100872, China
xwu@public3.bta.net.cn