PLS회귀를 이용한 포지셔닝맵의 구축

Building Positioning Map by PLS Regression

PLS회귀를 이용한 포지셔닝맵의 구축

Building Positioning Map by PLS Regression

이성근 • Yi, Seong Keun, 최지호 • Choi, Jiho, 이종호 • Lee, Jong-Ho

Partial Least Squares (PLS) 회귀 방법은 관찰치의 수보다 변수의 수가 더 많을 때 사용할 수 있는 다변량 분석기법이다. 뿐만 아니라 PLS회귀는 주성분을 추출할 때 반응변수와 설명변수를 동시에 고려하기 때문에 주성분회귀분석보다 예측력이 더 우월하다. 이 논문에서는 PLS방법으로 얻어진 주성분에 각 변수(속성)를 회귀시켜 각 변수의 벡터를 구하였으며, 각 주성분점수를 활용하여 관찰치(브랜드)들의 좌표를 구하여 각 관찰치들의 포지션을 지도상에 표시할 수 있도록 하였다. 얻어진 관찰치들의 포지션은 각 변수의 위치와 교차하여 해석할 수 있어 각 관찰치의 특성을 파악할 수 있으며, 각 설명변수들도 각 반응변수들과 어떻게 관계를 가질 수 있는가도 지도상에서 해석이 가능하다.

핵심주제어: PLS, Partial Least Squares, PLSR, PLS Regression, singular value decomposition, PCR, Principal Component Regression

이 성 근 | 성신여자대학교 사회과학대학 경영학과 부교수(yisk@sungshin.ac.kr), 주저자, 교신저자
최 지 호 | 전남대학교 경영학부 조교수(jihocool@chonnam.ac.kr)
이 종 호 | 고려대학교 경영대학 경영학과 조교수(jongholee@korea.ac.kr)

ISSN: 1229-456x (print)

ABSTRACT

Partial Least Squares regression(PLSR) proposed by Herman Wold in 1966 has been used as very valuable method to predict a set of response variables by a set of explanatory variables. PLSR is very useful for building a predictive model when variables are many and highly correlated. Multiple regression analysis also useful tool for building a prediction model. But it has much limitation when variables are many and highly correlated. In such cases, even though we can build a prediction model, it will fail to predict new data well (Tobias, 2007). Especially when the number of variables is much larger than the number of observations, the phenomenon of so-called 'overfitting' occurs. When the explanatory variables are highly correlated, one approach to overcome the problem is to remove the some of highly correlated explanatory variables. Another approach is to reduce the explanatory variables into small number of variables which have no correlations. The concept of PLS is to extract small numbers of latent variables which explain for highly correlated many variables. In that sense, PLS is a indirect modelling. But the way of extracting latent variables is different from the traditional method,

The superiority of PLSR to PCR(Principal Component Regression) is very well known. Ryan et. al (1999) showed empirically that PLSR is better than PCR in prediction the response variable. They compared three models with mediators and collinearity among the response variables, for example, regression, PCR, and PLSR. As the hypothesized conceptual model had moderators and collinearity in their study, the regression model was not germane to the research objective. Hence their focus was on the comparison of PCR, with PLSR. Even though the fact that there was a difference in estimating the coefficients between PCR and PLSR was very confusing, But prediction of PLSR was better than PCR.

Even though PLSR began in social sciences, it's uses are extended to the various fields like chemometrics (Westerhuis 1998; Wagon & Kowalski 1988; Geladi & Kowalski 1986) or sensory evaluation (Martens & Naes 1989), marketing (Abdi 2003; Chin et al. 2003; Graver, et al 2002; Ryan et al 1999; Fornell and Bookstein 1982; Japal 1982) and design (Han and Yang 2004). Interestingly, similarly to this research, Husson & Pages (2005) proposed the way of corresponding additional variables by the use of PLSR coefficients instead of the linear regression coefficients in Prefmap technique.

Huh and colleagues proposed several quantification methods using traditional multivariate data analysis techniques (Kim, 2000; Yang, 1998; Park and Huh 1996a, b; Han, 1995). The quantification methods proposed by them are endeavors to reduce the multivariate data with interrelationship and to represent or to plot them onto the low dimensional space. Projection pursuit stands for those methods. It aims to analyze the characteristics and structure of data through projecting the multivariate data onto the lower dimensional space and through analyzing the projection. In that sense, quantification method means a technique for building map in marketing.

The purpose of this research is to propose the algorithm for building positioning map by PLSR. The basis of the algorithm is a singular value decomposition. To derive the form of singular value decomposition, Lagrange multiplier method function was adopted. After components are extracted via singular value decomposition, the relationships between components and variables can be gotten by regressing variables on the components. The regression coefficients are the coordinates of the variables. Additionally we can get score vectors of components for observations from the same process. They are the coordinates of the observations. That is, The variables and observations can be positioned on the simple space generated by PLSR.

The quantification technique for PLS method gives us the better understanding of structure of variables and observations. The limitation of this study is the situation when there are more than 2 sets of data. In that case it is very to difficult to solve the Lagrange multiplier method function due to the many constraints in the equation. Thus we should consider another method of extracting the principal components due to the many constraints in the equation.

Key words: PLS, Partial Least Squares, PLSR, PLS Regression, singular value decomposition, PCR, Principal Component Regression

This paper is a part of the doctoral dissertation of the 1st author.
Yi, Seong Keun | 1st and Corresponding Author, Associate Professor, Dept. of Business Administration, Sungshin Women's Univ.
Choi, Jiho | Assistant Professor, School of Business Administration, Chonnam National Univ.
Lee, Jong-Ho | Assistant Professor, Dept. of Business Administration, Korea Univ.

Ⅰ. Backgrounds and Purposes

Partial Least Squares (PLS) proposed by Herman Wold in 1966 has been used as very valuable method to predict a set of response variables by a set of explanatory variables. Even though PLS began in social sciences, it's uses are expanded to various fields like chemometrics (Westerhuis 1998; Wagen & Kowalski 1987; Geladi & Kowalski 1986) or sensory evaluation (Martens & Naes 1989), marketing (Abdi 2003; Chin et al. 2003; Graber, et al 2002; Ryan et al 1999; Fornell and Bookstein 1982; Japal 1982) and design (Han and Yang 2004). Interestingly, similarly to this research, Husson & Pages (2005) proposed the way of corresponding additional variables by the use of PLS regression coefficients instead of the linear regression coefficients in Prefmap technique.

Basically PLS is a regression method in the sense that it deals the relationship between response variable and explanatory variables. But it is comparatively different method that it can be used when explanatory variables have collinearity, and when the number of observation is smaller than the number of explanatory variables. That is, univariate PLS, which deals one response variable has much similarity with traditional regression.

To avoid collinearity in PLS, it uses a kind of principal component analysis, which tries to reduce many mutually interrelated variables into small number of irrelevant variables. From that point of view, PLS can be compared to Principal Component Regression (PCR). The idea of PCR is similar to PLS. Both PCR and PLS consider collinearity or interrelationship among the explanatory variables. Whereas PCR considers only collinearity or interrelationship among the explanatory variables, PLS considers both response variables and explanatory variables when it determines the principle component. That is, PCR considers solely explanatory variables when it determines principle component, whereas, in PLS the information of response variables is considered when it determines principle component.

PLS is very similar to canonical correlation analysis(CCA) from the point of view that they deal the relationship of sets of variables and use a kind of principle component analysis (PCA).

Huh and his colleagues proposed several positioning map methods using traditional multivariate data analysis techniques (Kim 2000; Yang 1998; Park and Huh 1996a, b; Han 1995). They called it as 'quantification method'. The quantification methods proposed by them are endeavors to reduce the multivariate data with interrelationship and to represent or to plot them onto the lower dimensional space. Projection pursuit stands for those methods. It aims to analyze the characteristics and structure of data through projecting the multivariate data onto the lower dimensional space and through analyzing the projection. In that sense, quantification method means a technique for building map in marketing. Based on the above ideas, the purposes of this research is to propose an algorithm for building positioning map by partial least squares regression.

Ⅱ. Basic Ideas of the Study

1. Singular Value Decomposition

As told, the basis of the study is singular value decomposition (SVD). Let's consider the $n \times p$ data matrix X with rank r ($r < p$) (Huh 1995). X can be written as

$$X = UDV^t, \tag{2.1}$$

where $U = (u_1, u_2, \ldots u_r)$ and $V = (v_1, v_2, \ldots v_r)$ are the column orthogonal matrics of size $n \times r$ and $p \times r$

respectively($U^t U = I_r, V^t V = I_r$), and D is $r \times r$ diagonal matrix with singular value $\mu_1 \geq \mu_2 \cdots \geq \mu_r (>0)$ as its diagonal elements. The left singular vectors $U = (u_1, u_2, \ldots u_r)$ form an orthogonal basis for the columns of X in R^n and the right singular vectors $V = (v_1, v_2, \ldots v_r)$ form an orthogonal basis for the rows of X in R^p.

SVD has very close relationship with eigen system. Let's consider the data matrix $X^t X \ (p \times p)$ where X is $n \times p$ matrix. $X^t X$ can be written as

$$X^t X = VD^2 V^t. \tag{2.2}$$

We can find that the right singular vectors are eigen vectors of $X^t X$ and the squared singular values are eigen values, $\lambda_1, \lambda_2 \ldots \lambda_r$. So we can get eigen values and eigen vectors of $X^t X$ through SVD of X.

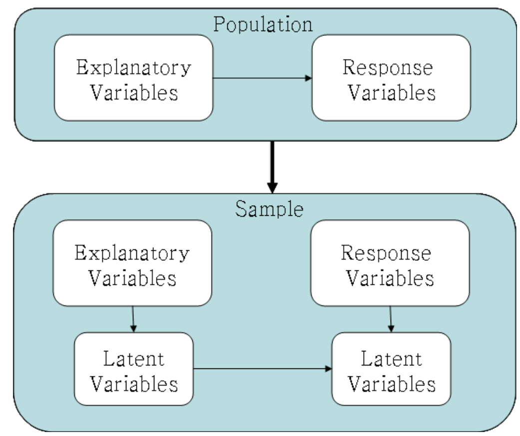
2. Partial Least Squares Regression

Partial least squares regression(PLSR) is very useful for building a predictive model when variables are many and highly correlated. Multiple regression analysis also useful tool for building a prediction model. But it has much limitation when variables are many and highly correlated. In such cases, even though we can build a prediction model, it will fail to predict new data well (Tobias 2007). Especially when the number of variables is much larger than the number of observations, the phenomenon of so-called 'overfitting' occurs.

When the explanatory variables are highly correlated, one approach to overcome the problem is to remove the some of highly correlated explanatory variables. Another approach is to reduce the explanatory variables into small number of variables which have no correlations.

The concept of PLS is to extract small numbers of latent variables which explain for highly correlated many variables. In that sense, PLS is a indirect modelling. But the way of extracting latent variables is different from the traditional method, It will be explained later.

〈FIGURE 2.1〉 Indirect Modelling



Source: Tobias, Randall D (2007).

PLS extracts the latent variables to maximize the relationship between the successive pairs of latent variables. So, it has interest in SVD of $X^t Y$. Originally PLS method means PLS regression. Usually we can get better predicted value through PLS regression than any other analysis. In PLS Y is a matrix of response variables and X, a matrix of explanatory variables. It menas that Y depends on X.

PLS proposed by Herman Wold in 1966, was improved by Naes and Martens (1985). Let's consider their original concept. Like PCR, PLS regression (PLSR) is a dependency model. For convenience, I will suggest such a case that X is a set of variables(or a matrix) and y is a single variable (or a vector). The X-variables and y-variables are scaled and centered, yielding X_0 and y_0. Then step 1 to 5 are performed for each component $a = 1, 2, \ldots, A_{\max}$ where $A_{\max}$ is the maximum number of PLSR component to be computed.

Step 1. Find weight vector $\widehat{w_a}$ by maximizing the covariance between the linear combination $X_{a-1}w_a$ and y under the constraint that $w_a^t w_a = 1$.

Step 2. Find factor scores, $\hat{t_k}$ as the projection of X_{a-1} on $\widehat{w_a}$, i.e. $\hat{t_a} = X_{a-1}\widehat{w_a}$

Step 3. Regress X_{a-1} on t_a to find the loading $\hat{p}_a^t$, i.e. $\hat{p_a} = X_{a-1}^t \ \hat{t_a} / \hat{t}_a^t \hat{t_a}$

Step 4. Regress y_{a-1} on t_a to find the loading $\hat{q_a}$, i.e. $\hat{q_a} = y_{a-1} \hat{t_a} / \hat{t}_a^t \hat{t_a}$

Step 5. Subtract $\hat{t_a} \hat{p}_a^t$ from X_{a-1} and call the new matrix X_a.

Subtract $\hat{t_a} \hat{q_a}$ from y_{a-1} and call the new matrix y_a.

After repeating the above steps until the maximum number of component, $A_{\max}$.

The superiority of PLSR to PCR is very well known. Ryan et.al(1999) showed empirically that PLSR is better than PCR in prediction the response variable. They compared three models with mediators and collinearity among the response variables, for example, regression, PCR, and PLSR. As the hypothesized conceptual model had moderators and collinearity in their study, the regression model was not germane to the research objective. Hence their focus was on the comparison of PCR, with PLSR. Even though the fact that there was a difference in estimating the coefficients between PCR and PLSR was very confusing, But prediction of PLSR was better than PCR.

III. Suggesting Algorithm and Numerical Example

1. Basic Concept

As discussed earlier, the purpose of PLS regression is a prediction of response variable(s). Let's consider data matrix K with p explanatory variables, q response variables and n observations. Data matrix K consists of $X(n \times p)$ and $Y(n \times q)$. Here, I assume that data matrix X and Y are scaled and centered, but no such transformation mandatory (Huh et al 2007; Yi 2007).

The aim of PLS regression is to find linear combination of p-explanatory variables (X) and q-response variables (Y) which maximizes the covariance between the projections of each sets of variables. The problem of maximization can be written as

$$\text{maximize (w.r.t. } a \text{ and } b\text{) } \operatorname{Cov}(Xa, Yb) \tag{3.1}$$
$$\text{subject to } a^t a = b^t b = 1,$$

where Xa and Yb are projections of each data matrix X and data matrix Y (Huh, Lee, and Yi, 2007).

As the covariance of (3.1) is dependent on both direction and norm of a and b, two constraints $a^t a = 1$ and $b^t b = 1$ are considered. Lagrangian function can be used to get the solution of (3.1) under the constraints. The function L is defined as

$$L(a, b, \lambda_1, \lambda_2) = a^t X^t Yb - \lambda_1 (a^t a - 1) - \lambda_2 (b^t b - 1) \tag{3.2}$$
$$\text{subject to } a^t a = 1 \text{ and } b^t b = 1 .$$

By setting the partial differential of L to 0_p and 0_q, (3.3) and (3.4) are obtained.

$$X^t Yb - 2\lambda_1 a = 0_p, \tag{3.3}$$

$$Y^t Xa - 2\lambda_2 b = 0_q \tag{3.4}$$

By solving the simultaneous equations of (3.3) and (3.4), with respect to a, b is eliminated, Consequently, we have

$$X^t YY^t Xa = 4\lambda_1 \lambda_2 a . \tag{3.5}$$

Here, the solution of a is an eigen vector of $p \times p$ non-negative matrix $X^t\ Y Y^t X$. In the same manner, the solution of b is an eigen vector of $q \times q$ non-negative matrix $Y^t X X^t\ Y$. Therefore, both a and b can be obtained from SVD (singular value decomposition) of $p \times q$ matrix $X^t\ Y$. That is, by the use of (3.5), $X^t\ Y Y^t X = U D_\mu V^t\ V D_\mu U^t$ is obtained. Since $V^t V = I_q$, $X^t\ Y Y^t X$ is equal to $U D_{\mu^2} U^t$. So, by decomposing $X^t\ Y Y^t X$, eigen value $\lambda = \mu^2$ and each eigen vector of matrix $X^t\ Y Y^t X$ is obtained.

Similarly, matrix $Y^t X X^t\ Y$ is decomposed into $V D_{\mu^2} V^t$, and eigen value $\lambda = \mu^2$ and each corresponding eigen vector of matrix $Y^t X X^t\ Y$ is obtained. Thus, we can find that weight vector a of matrix X is u_1 and weight vector b of matrix Y is v_1 from

$$X^t\ Y = U D_\mu V^t, U = (u_1, u_2, \cdots), V = (v_1, v_2, \cdots), \quad \mu_1 \geq \mu_2 \geq \cdots, \tag{3.6}$$

where $U^t U = V^t V = I_q$ and D_μ is a matrix with singular values $\mu_1 \geq \mu_2 \geq \cdots$ on the diagonal, the columns $u_1, u_2, \cdots$ of U are left singular vectors, and the columns $v_1, v_2, \cdots$ of V are right eigen vectors. Usually eigenvalues λ_i is referred to μ_i^2.

Consider the case that Y is a vector ($=y$). In PLS method, b is obtained as $X^t y \,/\, \| X^t y \|$, when we consider the constraint ($b^t b = 1$) in (3.1). Accordingly SVD for Y is not needed.

2. Positioning Algorithm for PLS regression

Let's consider data matrix K with p explanatory variables, q response variables and n observations. Xa will be transcribed as s ($n \times 1$) and call it as score vector. By regressing Y on s, Y fit, $\hat{Y}$ can be obtained as

$$\hat{Y} = s\,(s^t s)^{-1} s^t\ Y = s\, g_Y^t,$$
$$\text{where } g_Y^t = (s^t s)^{-1} s^t\ Y. \tag{3.7}$$

g_Y ($q \times 1$), regression coefficients of s, can be called Y-loading vector. As shown in (3.7), PLS regression is a simple linear regression method in which Y is regressed on an explanatory variable s.

When we determine coefficient vector b (in $s = Xb$), we should consider Y as well as X simultaneously. Determining the regression coefficient g_Y^t is very complicated, since $\hat{Y} = A\ Y$, where $A = A(Y)$ or $s\,(s^t s)^{-1} s^t$. Accordingly, distribution of $\hat{Y}$ can not be obtained easily in PLS regression, contrary to linear regression.

Above procedure is the first step in PLS regression. We can identify that rank of transformational matrix A of Y is 1. To obtain the improved fit, the rank of A can be increased by the use of technique which will be shown in the quantification step. For convenience, I will suggest one more step in this algorithm. I will use the following notations for convenience.

Notations

- K : data matrix with several sub data matrix
- X, Y : sub data matrix
- a, b : weight vectors obtained from SVD
- s, t : score vectors of sub data matrix X, Y
- g_X : loading vector for sub data matrix X
- g_Y : loading vector for sub data matrix Y
- $\hat{X}$: predicted value of sub matrix X
- $\hat{Y}$: predicted value of sub matrix Y
- number in subscript : PLS cycle

For the further steps, I will denote $s \rightarrow s_1$, $g_Y \rightarrow g_{1,Y}$,

$X \to X_1$, $\hat{X} \to \widehat{X_1}$ and $\hat{Y} \to \widehat{Y_1}$. Following steps can be put in order for quantification of PLS regression.

- Data are centered and scaled

Cycle 1

Step 1 : Find weight vectors and score vectors

Find a_1 and b_1 in the manner of maximizing (3.1) under the these constraints $a^t a = 1$, $b^t b = 1$. Accordingly, we can obtain score vectors ($X_1 a_1 = s_1$ and $Y_1 b_1 = t_1$) of data matrix X_1 and Y_1.

Step 2 : Find loading vectors

Obtain $\widehat{X_1}$ and $\widehat{Y_1}$. $\widehat{X_1}$ can be obtained by regressing X_1 on s_1 and $\widehat{Y_1}$ can be obtained by regressing Y_1 on s_1. Thus, $\widehat{Y_1}$ is $s_1 (s_1^t s_1)^{-1} s_1^t Y_1$ $(= s_1 g_{1.Y}^t)$, where $g_{1.Y}^t = (s_1^t s_1)^{-1} s_1^t Y_1$ and $\widehat{X_1}$ is $s_1 (s_1^t s_1)^{-1} s_1^t X_1$ $(=s_1 g_{1.X}^t)$. where $g_{1.Y}^t = (s_1^t s_1)^{-1} s_1^t X_1$.

If $X_1^t s_1 (s_1^t s_1)^{-1}$ is denoted to $g_{1.X}$, then $\widehat{X_1}$ become $s_1 g_{1.X}^t$. Here $g_{1.X}^* (p \times 1)$ $(= X_1^t s_1 / \| s_1 \|)$ is called X-loading vector. Similarly, $g_{1.Y}$ $(q \times 1)$ can be called Y-loading vector.

Step 3 : Deflate the data

Deflate X_1 and Y_1 with a following manner.

$$Y_2 = Y_1 - \hat{Y}_1,\ X_2 = X_1 - \hat{X}_1 .$$

Cycle 2

Step 4 : Finding weight vectors and score vectors

Find a_2 and b_2 in the manner of maximizing Cov($X_2 a_2$, $Y_2 b_2$) under the constraints $a_2^t a_2 = 1$ and $b_2^t b_2 = 1$. Compute new score vector s_2 and t_2.

Step 5 : Find loading vectors

Obtain $\widehat{X_2}$ and $\widehat{Y_2}$. $\widehat{X_2}$ can be obtained by regressing X_2 on s_2 and $\widehat{Y_2}$ can be obtained by regressing Y_2 on s_2, where $s_2 = X_2 a_2$, $t_2 = Y_2 b_2$.

$$\widehat{Y_2} = s_2 (s_2^t s_2)^{-1} s_2^t Y_2 = s_2 g_{2.Y}^t,$$
$$\widehat{X_2} = s_2 (s_2^t s_2)^{-1} s_2^t X_2 = s_2 g_{2.X}^t$$

Consequently, $\hat{Y}$ and $\hat{X}$ can be expressed as follows.

$$\begin{aligned}
\hat{Y} &= \widehat{Y_1} + \widehat{Y_2} \\
&= s_1 (s_1^t s_1)^{-1} s_1^t Y_1 + s_2 (s_2^t s_2)^{-1} s_2^t Y_2 \\
&= s_1 g_{1.Y}^t + s_2 g_{2.Y}^t, \\
\hat{X} &= \widehat{X_1} + \widehat{X_2} \\
&= s_1 (s_1^t s_1)^{-1} s_1^t X_1 + s_2 (s_2^t s_2)^{-1} s_2^t X_2 \\
&= s_1 g_{1.X}^t + s_2 g_{2.X}^t,
\end{aligned}$$

where $g_{2.Y}^t = (s_2^t s_2)^{-1} s_2^t Y_2$

and $g_{2.X}^t = (s_2^t s_2)^{-1} s_2^t X_2$.

By the use of above procedure, the prediction value of subject $y^* (q \times 1)$ that has $x^* (p \times 1)$ is possible. The procedure extends iteratively in a natural way to give r (r =1,2,3,⋯) number of components of $\hat{X}$ and $\hat{Y}$. To determine the number of components which will be included in regression model, cross validation technique is usually used.

The focus of this paper lies in the suggesting positioning method using PLS regression. Consider multivariate data matrix K which consists of data matrix X with p-explanatory variables and data matrix Y with q-response variables again. I assume that data matrix X and data matrix Y are scaled and centered. According to PLS regression, score vectors of X, s_1, s_2 $(=n \times 1)$ are orthogonal and they can generate the base of projection space.

For positioning of multivariate data matrix (X, Y) by PLS in the reduced space, determination of the coordinates is needed. In step 2 and step 5 of the algorithm, as previously showed, we obtained X-loading vectors and Y-loading

〈TABLE 3.1〉 Quantification formulas of PLS regression for columns (variables)

	Dimension 1	Dimension 2
X variables	$x_j^t s_1^*$	$x_j^t s_2^*$
Y variables	$y_k^t s_1^*$	$y_k^t s_2^*$

Note : $s_1^* = s_1/\parallel s_1 \parallel$, $s_2^* = s_2/\parallel s_2 \parallel$

〈TABLE 3.2〉 Quantification formulas of PLS regression for rows (observations)

	Dimension 1	Dimension 2
Observations in data matrix X	$X_1 a_1 = s_1$	$X_2 a_2 = s_2$
Observations in data matrix Y	$Y_1 b_1 = t_1$	$Y_2 b_2 = t_2$

vectors. Each loading vectors for $X(=n\times p)$ variables and $Y(n\times q)$ variables can be used as coordinates. Thus, columns $x_j(j=1,2,...p)$ of data matrix X can be pointed on the linear space P_j : $(x_j^t s_1^*,\ x_j^t s_2^* \cdots)$ generated by s_1, s_2, ・・・. And columns $y_k(k=1,2,...q)$ of data matrix Y can be pointed on the linear space Q_k : $(y_k^t s_1^*,\ y_j^k s_2^* \cdots)$ generated by s_1, s_2, ・・・. Here s_1^* is $s_1^* = s_1/\parallel s_1 \parallel$ and s_2^* is $s_2^* = s_2/\parallel s_2 \parallel$. The coordinates of each columns for dimension 1 and for dimension 2 are suggested in Table 3.1 and Table 3.2

3. Numerical Example

Data description

The data shown as an example here are the survey results of the automobile market in China. I am interested in how the property of automobile has an effect on the consumer's attitude toward brand. Thus I considered the data for property evaluation of the automobile and the data for attitude toward brand which are collected from the survey done in 2006.

Thus, I consider the data matrix K with thirty six variables and fifty observations (companies). Data matrix K consists of two sets of variables denoted by $X(=50\times 34)$ and $Y(=50\times 2)$. Here, X is a data set for consumers' evaluation of automobile's property for companies. And Y is a data set for consumers' attitude toward brand. Here, data set X and Y are collected in a seven point scale (from point 1 to point 7) and they are scaled and centered for the analysis. The brands and properties evaluated are listed in Table 3.1 and Table 3.2. Attitude toward brand used as a data set Y are 'overall satisfaction $(= Y1)$' and 'repurchase intention $(=Y2)$'.

Interpretation of the result

Loading vectors of X, Y variables are listed in Table 3.5 and in Table 3.6, and score vectors are listed in Table 3.7. Quantification plots are showed in Figure 3.1 based on the Table 3.5 and Table 3.6.

Two components were extracted for convenience in this analysis. The total amount of the variance which was explained by two components was 56.3%. The first component explained the variance by 37.2% and the second component did 19.0%.

The X-variables are divided into two groups on the

〈TABLE 3.3〉 Automobile brands surveyed in China

1. Beijing Hyundai 2. Beijing Jeep 3. Changan Ford 4. Changan Suzuki 5. Dongfeng Citroen 6. Dongfeng honda 7. Dongfeng Nissan 8.DYK 9. Southeast Motor 10.Faw Hainan Mazda 11. Faw Mazda 12.Faw-VW 13. Guangzhou Honda 14. Guangzhou Toyota 15. Geely 16. Nanjing Fiat 17. Chery 18.Shanghai GM 19. SVE 20. Tianjin Faw 21. Faw Toyota 22. Korean Hyundai 23. Korean Kia 24. Hafei Motor 25. Changhe Suzuki 26. Changan Motor 27. Biyadi 28. Jiangnan Auto 29. Faw Huali 30. Jilin Tongtian 31. Dongfeng Liuzhou 32. Nanjing Motor 33. Dongfeng Peugeot 34. SGM Wuling 35. Shanghai Maple 36. Beijing Benz 37. Huachen BMW 38. Huachen Motor 39. Faw Motor 40. Changcheng Auto 41. Changfeng Auto 42. Jiangling Auto 43. Zhengzhou Nissan 44. Jiao Auto 45. Huatai Hyundai 46. Beijing Futon 47. Beijing Auto 48. Jianghuai Auto 49. Baolong Auto 50. Mercedes-Benz

〈TABLE 3.4〉 Property list of automobile brands

1. proper engine displacement/ power 2. engine type (v6,diesel engine, etc.) 3. good acceleration 4. good performance in cross country running. 5. stability at steering 6. convenience for parking 7. durability of the whole 8. type of drive (two-wheel/four-wheel drive) 9. gear type (manual/auto) 10. overall exterior styling 11. overall interior 12. broad vision 13. car size 14. convenience to get in and out 15. convenience to load and unload cargoes 16. space of the front seats 17. space of the second row 18. overall quietness 19. standard features 20. price 21. scope of quality guarantee 22. future trading price 23. efficiency of fuel 24. efficiency of maintenance 25. manufacturer impression 26. place of origin 27. availability of parts 28. cargo capacity 29. guard against theft 30. overall safety 31. environmental protection 32. sales service 33. after-sales service 34. lead time

direction. The variables of the first group gather around variable 30 (overall safety). The first group consist of variable 30 (overall safety), variable 25 (manufacturer's impression), variable 13 (car size), variable 10 (overall exterior styling) and so forth. The variables of the second group gather around variable 22 (future trading price). The second group consist of variable 22 (future trading price), variable 20 (price), variable 15 (convenience to load and unload cargoes), variable 29 (guard against theft) and so forth. We can interpret that the first group is on 'the basic performance or the function of the automobile' and the second one is on 'the additional value of the automobile'.

The observations (brands) are dense around the second axis and scattered along the first axis. We can interpret that there is no substantial difference in the second axis and some difference in the first axis among the observations (brands). That is, the difference among the brands occur only in the first axis.

By the use of the loading vectors and score vectors, variables and observations are plotted onto the space generated by each score vectors. They are Figure 3.1 and Figure 3.2. As shown in Figure 3.1, variable 7 in X-variables ('durability of the whole') has same direction with variables 1 ('overall satisfaction') of Y-variables. It means that 'durability of the whole' has a relationship with 'overall satisfaction'. Similarly variable 18 ('overall quietness') and variable 10 ('overall exterior styling') of X variables are very close to variable 2 ('repurchase intention') of Y variables.

〈TABLE 3.5〉 Loading vectors of X variables

variables	Loading vectors of X variables	
	dimension 1	dimension 2
X1	-4.582	-0.726
X2	-4.013	1.195
X3	-5.303	-0.001
X4	-3.256	-2.100
X5	-5.859	-1.520
X6	-1.046	-3.036
X7	-3.703	3.934
X8	-4.628	-2.451
X9	-0.177	0.614
X10	-3.890	1.458
X11	-4.666	-0.967
X12	-3.804	1.291
X13	-4.183	-0.545
X14	-3.747	-3.169
X15	-2.611	-4.947
X16	-5.642	-1.613
X17	-1.802	1.425
X18	-5.230	1.758
X19	-3.796	-1.001
X20	0.436	-3.327
X21	-3.799	-2.655
X22	-1.648	-5.133
X23	-1.693	-3.586
X24	-3.694	0.636
X25	-5.302	-0.830
X26	-4.799	-2.719
X27	-2.566	-4.022
X28	-4.223	-2.094
X29	-4.316	-1.756
X30	-5.704	1.072
X31	-2.245	-3.852
X32	-5.429	-2.204
X33	-3.836	-2.860
X34	-4.817	-3.710

We can interpret the plots in Figure 3.1 and Figure 3.2 jointly. To the direction of 'overall satisfaction', observation 49 ('Baolong auto') and observation 36 ('Beijing Benz') locate. Observation 37 ('Huachen BMW') and 43 ('Zhengzhou Nissan) have very close relationship ('repurchase intention'). That is to say, we can infer that those brands are well evaluated in the 'attitude toward brand'.

We can combine the plots of X-variables with and X-observations. Brand 46 (Beijing Futon), brand 47 (Beijing Auto) and brand 30 (Jilin Tongtian) have the direction with the second group of the X-variables. It can be interpreted that Brand 46 is evaluated most positively in the second

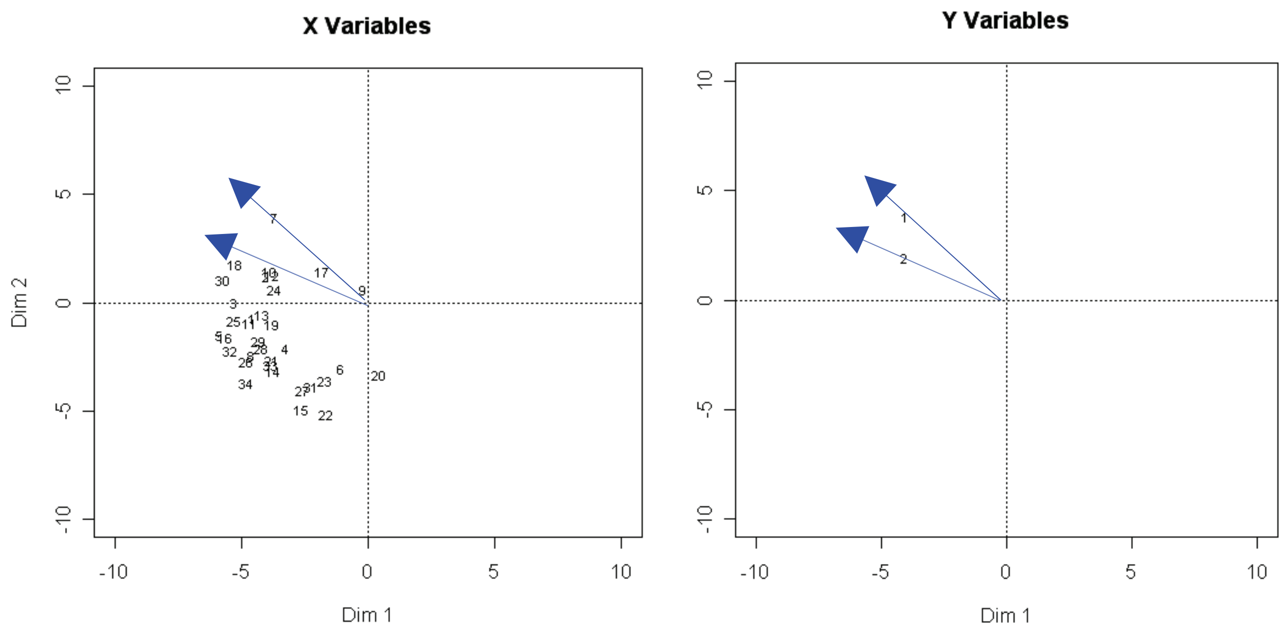
〈TABLE 3.6〉 The loading vectors of Y variables

variables	Loading vectors of Y variables	
	dimension 1	dimension 2
Y1	-4.018	3.843
Y2	-4.061	1.926

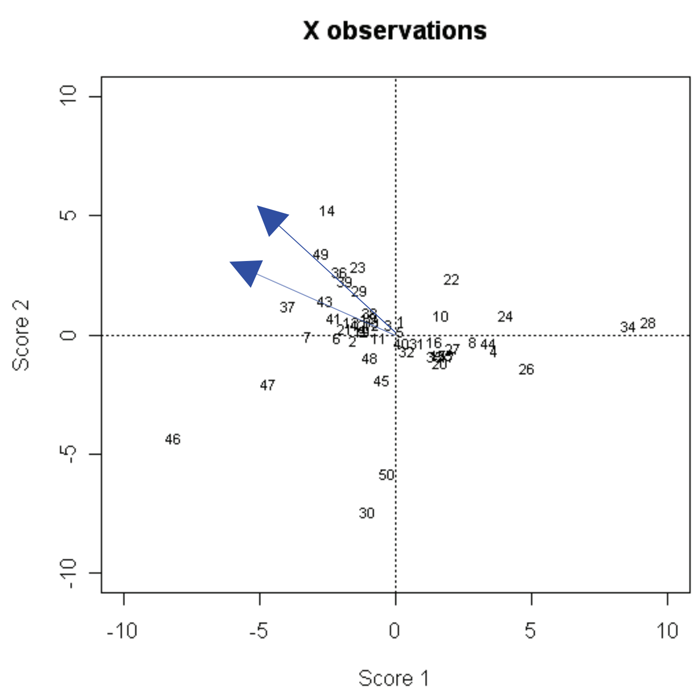
〈TABL 3.7〉The score vectors of X and Y

brands	Score vectors of X		Score vectors of Y	
	score 1	score 2	score 1	score 2
1	0.2573	0.6153	-0.1716	0.1950
2	-1.5322	-0.2144	-0.7644	0.4854
3	-0.2506	0.4663	-1.1950	1.0885
4	3.6472	-0.6619	0.8696	0.1628
5	0.1907	0.1543	-0.0462	0.2253
6	-2.1338	-0.0873	-0.1893	-0.5392
7	-3.2126	-0.0604	-0.7998	-0.1804
8	2.8621	-0.2684	0.1691	0.7810
9	-1.1962	0.1422	0.5466	-0.8871
10	1.6745	0.8386	0.2413	0.1749
11	-0.5950	-0.0872	-0.6022	0.2459
12	-0.8816	0.4228	-0.9075	0.4908
13	-1.5963	0.5584	-0.7998	0.2379
14	-2.4841	5.2791	-2.9721	2.0106
15	1.6042	-0.8362	1.6055	-0.9565
16	1.4725	-0.3022	0.4567	-0.0131
17	1.9291	-0.9091	0.8696	-0.2819
18	-0.9682	0.7146	-0.6921	0.3326
19	-1.1885	0.1760	-0.6921	0.2756
20	1.6584	-1.1367	0.2591	0.2862
21	-1.8311	0.2621	-0.6022	-0.0740
22	2.1013	2.4058	-1.1051	1.5139
23	-1.3366	2.8859	0.1691	-0.3057
24	4.0785	0.8092	1.4624	-0.4791
25	1.8296	-0.8445	0.3490	0.1472
26	4.8526	-1.3732	1.4978	-0.0479
27	2.1274	-0.5577	0.9050	0.0003
28	9.3323	0.5278	2.2514	0.6364
29	-1.3013	1.9018	-0.8366	0.8439
30	-1.0190	-7.4022	1.1408	-2.3552
31	0.8145	-0.3278	-0.0639	0.2714
32	0.4360	-0.6694	0.1337	-0.0778
33	-0.8699	0.5462	-0.2615	0.0866
34	8.5661	0.3622	2.9710	-1.0283
35	1.4476	-0.8746	1.7131	-1.0649
36	-2.0491	2.6768	-1.2658	0.1614
37	-3.9161	1.2294	-1.2127	0.0244
38	-0.9434	0.9541	-0.7821	0.5223
39	-1.8564	2.2781	-2.0209	1.4468
40	0.2288	-0.3174	0.4744	-0.2196
41	-2.2228	0.7235	-0.5668	0.0555
42	-1.3820	0.4531	1.0672	-1.3900
43	-2.5655	1.4262	-0.2615	-0.3522
44	3.4582	-0.3356	1.4978	-0.4088
45	-0.4757	-1.8626	1.1203	-0.8091
46	-8.1641	-4.3247	-0.2084	-1.4549
47	-4.6658	-2.0445	-1.2850	0.1292
48	-0.9320	-0.9416	0.6543	-0.8866
49	-2.7140	3.4292	-2.8113	2.2295
50	-0.2852	-5.7993	0.6911	-1.2488

〈FIGURE 3.1〉 Plots of variables by PLS regression



〈FIGURE 3.2〉 Plots of observations by PLS regression



group of variables. But as the variables of the second group has no relationship with the 'attitude toward brand', it seems that the good evaluation of those brands will not be associated with the direct selling. On the other hand, brand

37 (Huachen BMW) has a same direction with the first group of variables. Thus it seems that brand can get good performance in the market.

On the contrary, brand 34 (SGM Wuling) and brand 28 (Jiangnan Auto) have a opposite direction to the other variables. It seems that they are badly evaluated in the properties of X-variables. Brand 44 (Jiao Auto) and brand 26 (Changan Motor) have similar position with brand 34 and brand 28.

To get the visual image I suggested in this paper, I used 'R' language.

Ⅳ. Summary and Discussions

The purpose of this research is to propose the positioning algorithm for PLS regression. In this study I proposed how to position the variables and the observations onto the simple space by PLS regression. The basis of the algorithm is in the singular value decomposition. But the problem exists in the way of deriving the singular value composition.

To derive the form of singular value decomposition, Lagrange multiplier method function was adopted. After components are extracted via singular value decomposition, the relationships between components and variables can be derived by regressing variables on the components. The regression coefficients are the coordinates of the variables. Additionally we can get score vectors of components for observations. They are the coordinates of the observations. Based on the coordinates, the variables and observations can be positioned on the simple space generated by PLS regression.

The quantification technique for PLS method gives us the better understanding of structure of variables and observations. Especially when there are so many sets of variables, quantification technique proposed here is very useful. As we mentioned above, the key idea of this algorithm lies in the way of building the singular value composition format. When there are two sets of data, using Lagrange multiplier method function may be a good way of building the singular value composition format. But, given the over 3 sets of data, it is very to difficult to solve the Lagrange multiplier method function due to the many constraints in the equation.

Let's consider 3 sets of variables $X(=n\times p)$, $Y(=n\times q)$, and $Z(=n\times r)$. Let denote Xa, Yb, and Zc be the projections of each data matrix X, Y, and Z. Unlike objective function suggested in the case of two sets of variables, we have to use the constraints $a^t a=1, b^t b=1,$ and $c^t c=1$ for obtaining solution. In this case, the method we used in the two data sets case has problem in solving the problem. Of course, alternatively, the constraint $a^t a+b^t b+c^t c=3$ can be used for the simple process. Strictly we can not be sure that it should be a correct way of solving problem. For that reason, it is very needful to find a way of solving the problem in the case of many data sets.

〈최초투고일: 2009년 10월 26일〉
〈수정일: 1차: 2009년 12월 28일〉
〈게재확정일: 2011년 6월 13일〉

References

Abdi, Herve (2003), "Partial Least Squares (PLS), Regression," in Lewis -Beck M., Bryman, A., Futing T.(Eds), *Encyclopedia of Social Sciences Research Methods,* Thousand Oaks (CA), Sage, 1-7

Chin, Wynne W., Marcolin, Barbara L. , and Newsted, Peter R. (2003), "A Partial Squares Latent Modelling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion/ Adoption Study," *Information Systems Research,* 14(2), 189-217

de Jong, S. (1993), "SIMPLS: an alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 109-119

Fornell, C., and Bookstein, F. (1982). "Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory," *Journal of Marketing Research*, 19, 440-452

Garthwaite, Paul H (1994), "An Interpretation of Partial Least Square," *Journal of the American Statistical Society*, 89 (425), 122-127

Graber, Stephane, Czllar, Sandor and Denis, Jean-Emile (2002), "Using Partial Least Squares Regression in Marketing Research," *unpublished working paper*, University of Geneva

Geladi, P, and Kowalski, B. (1986), "Partial Least Squares Rregression: A Tutorial," *Analytica Chimica Acta*, 185, 1-17.

Han, Sang-Tae (1995). *Quantification Approach to Ranked Data Analysis,* Doctoral Dissertation, Korea University

Han, Sung H. and Yang, Huichul (2004), "Screening Important Design Variables for Building a Usability Model: Genetic Algorithm-based Partial Least-Squares Approach, "*International Journal of Industrial Ergonomics*, 33, 159-171

Husson, Francois and Pages, Jerome (2005), "Scatter Plot and Additional Variables," *Journal of Applied Statistics*, 32(4), 341-349

Huh, Myung-Hoe (1999), *Quantification Methods for Multivariate Data*, Seoul: Free Academy

Huh, Myung-Hoe (1999), *Understanding Quantification Methods*, Seoul: Free Academy

Huh, Myung-Hoe (1995), *Understanding and Computation the Matrix*, Seoul : Free Academy

Huh, Myung-Hoe, Lee, Yonggoo and Yi, Seong Keun (2007), "Visualizing (X,Y) Data by Partial Least Squares Method," *Korean Journal of Applied Statistics*, 20(2), 345-355

Helland, I. (2005), "Partial Least Squares Regression," *The Encyclopedia of Statistical Sciences*, Second Edition (edited by Kotz). 5957-5962.

Japal, Harsharanjeet S., (1982), "Multicollinearity in Structural Equation Models with Unobservable Variables," *Journal of Marketing Research*, 19(Nov.), 431-439

Kim, Mi-Kyung (2000), *Low Dimensional K-Means Clustering*, Doctoral Dissertation, Korea University

Martens, H. and Naes, T (1989), *Multivariate Calibration*, New York, John Wiley

Naes, T. and Martens, H. (1985), Comparison of Prediction Methods for Multicollinear Data, *Communications in Statistics - Simulation and Computations*, 14, 545-576.

Park, Mira and Huh, Myung-Hoe (1996a), "Canonical Correlation Biplot," *The Korean Communications in Statistics*, 3(1), 11-19

Park, Mira and Huh, Myung-Hoe (1996b), "Quantification Plots for Several Sets of Variables," *The Journal of Korean Statistical Society*, 25(4), 589-601

Ryan, Michael J., Rayner, Robert and Morrison, Andy (1999), "Diagnosing Customer Loyalty Drivers,"*Marketing Research*, Summer, 19-26

Tobias, Randall D. (2007), An Introduction to Partial Least Squares Regression, http://support.sas.com/techsup/technote/ts509.pdf

Wagon, L. E., and Kowalski, B. R. (1988), "A Multiblock Partial Least Squares Algorithm for Investigating Complex Chemical Systems," *Journal of Chemometrics*, 3, 3.

Wegelin, Jacob, A. (2000), *A Survey of Partial Least Square (PLS) Methods, with Emphasis on the Two-Block Case*, Technical Report, Department of Statistics, University of Washington.

Westerhuis, Johan A., Kourti, Theodora and Macgregor, John F. (1998), Analysis of Multiblock and Hierarchical PCA and PLS Models," *Journal of Chemometrics*, 12, 301-321

Wold, H.(1966), *Estimation of Principal Components and Related Models by Iterative Least Squares. in Multivariate Analysis* (ed. Krishnaiah P. R.) 391-420, Academic Press, New York.

Yang, Kyung-Sook (1998), *Correspondence Analysis Specific to Certain Types of Categorical Data*, Doctoral Dissertation, Korea University

Yi, Seong Keun (2007), *Quantification Method for Partial Least Squares and Its Generalization*, Doctoral Dissertation, Korea University