The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.1.225
25*(1), 225–237*

Generalization of Quantification for PLS Correlation

Seong Keun Yi[1] · Myung-Hoe Huh[2]

[1]Department of Business Administration, Sungshin Women's University
[2]Department of Statistics, Korea University

(Received September 23, 2011; Revised November 1, 2011; Accepted November 1, 2011)

Abstract

This study proposes a quantification algorithm for a PLS method with several sets of variables. We called the quantification method for PLS with more than 2 sets of data a generalization. The basis of the quantification for PLS method is singular value decomposition. To derive the form of singular value decomposition in the data with more than 2 sets more easily, we used the constraint, $a^ta + b^tb + c^tc = 3$ not $a^ta = 1$, $b^tb = 1$, and $c^tc = 1$, for instance, in the case of 3 data sets. However, to prove that there is no difference, we showed it by the use of 2 data sets case because it is very complicate to prove with 3 data sets. The keys of the study are how to form the singular value decomposition and how to get the coordinates for the plots of variables and observations.

Keywords: Partial Least Squares(PLS), generalization of quantification for PLS correlation.

1. Motivation, Problem and Concept of Generalization

Partial Least Squares(PLS) has been a very useful tool in the reduction of data when the number of observation is smaller than the number of variables (Helland, 2005). There are limited studies that compare the quantification methods of PLS with multivariate data analysis. Quantification is a method that tries to reduce and to visualize multivariate data into the lower dimensional space. Huh and his colleagues have tried this kind of endeavor (Huh *et al.*, 2007; Huh, 1999; Park and Huh, 1996; Han, 1995).

Based on the above mentioned quantification idea, we suggested how to quantify partial least squares with 2 sets of data (Huh *et al.*, 2007). We are now interested in quantification of PLS data with more than 2 sets. We will call it as a generalization of quantification for PLS correlation.

The PLS correlation can be extended to three or more sets of variables. Thus, we 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. On this occasion, the solution problem will be as follows. Unlike constraint suggested in the case of two sets of variables (Huh *et al.*, 2007), we use the constraint,

This study is a part of the first author's doctoral dissertation.
[1]Corresponding author: Associate Professor, 249-1, Dongseon-dong 3 ga, Seongbuk-gu, Department of Business Administration, Sungshin Women's University, Seoul 136-742, Korea. E-mail: yisk@sungshin.ac.kr

$a^ta + b^tb + c^tc = 3$ to get the solution more easily (Yi, 2007). The constraint $a^ta = 1$, $b^tb = 1$, and $c^tc = 1$ is a more natural form on this objective function; however, a simpler constraint is considered to solve the objective function more easily.

2. Approach and Solution

To show that it is not problematic for this simpler constraint to solve the objective function, we are going to compare a natural form of constraint ($a^ta = 1$, and $b^tb = 1$), with a simpler form of constraint ($a^ta + b^tb = 2$), in the case of two sets of variables for convenience. Let us consider following objective function.

$$\begin{aligned} &\text{maximize (w.r.t. } a, b, \text{ and } c) \;\; \mathrm{Cov}(Xa, Yb) + \mathrm{Cov}(Xa, Zc) + \mathrm{Cov}(Yb, Zc) \\ &\qquad \text{subject to } a^ta + b^tb + c^tc = 3. \end{aligned} \tag{2.1}$$

Lagrangian function can be used to get the solution of (2.1) under the constraints. Let us define the function L as

$$L(a, b, \lambda) = a^tX^tYb + b^tY^tZc + a^tX^tZc - \lambda(a^ta + b^tb + c^tc - 3). \tag{2.2}$$

By setting partial derivatives of (2.2) to 0_p, 0_q, 0_r respectively, we obtain

$$\frac{\partial L}{\partial a} = X^tYb + X^tZc - 2\lambda a = 0_p, \tag{2.3}$$

$$\frac{\partial L}{\partial b} = Y^tXa + Y^tZc - 2\lambda b = 0_q, \tag{2.4}$$

$$\frac{\partial L}{\partial c} = Z^tXa + Z^tYb - 2\lambda c = 0_r. \tag{2.5}$$

It is very complicated to solve the simultaneous equations of (2.3), (2.4) and (2.5). Let us define matrix E and D as follows.

$$E = \begin{bmatrix} X^tX & X^tY & X^tZ \\ Y^tX & Y^tY & Y^tZ \\ Z^tX & Z^tY & Z^tZ \end{bmatrix}, \quad D = \begin{bmatrix} X^tX & & \\ & Y^tY & \\ & & Z^tZ \end{bmatrix}, \quad \lambda = \text{constant}, \quad v = \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

where $v^tv = 3$ ($\because a^ta + b^tb + c^tc = 3$).

By the use of (2.3), (2.4) and (2.5), we consider following equation to derive eigensystem.

$$(E - D)v = 2\lambda v. \tag{2.6}$$

If we denote M for $(E - D)$, we obtain (2.7).

$$Mv = 2\lambda v, \tag{2.7}$$

where $M = \begin{bmatrix} 0 & X^tY & X^tZ \\ Y^tX & 0 & Y^tZ \\ Z^tX & Z^tY & 0 \end{bmatrix}$.

To find weight vectors, we can use eigenvalue decomposition. By the use of SVD of M, we can obtain eigenvalues and eigenvectors ($= v$). When M is used for obtaining meaningful eigenvalues and eigenvectors, matrix M should be positive definite; however, in this case we cannot guarantee that matrix M will be positive definite.

All eigenvalues are not necessarily positive after we obtain the eigenvalues by SVD. At least one positive eigenvalue corresponding to eigenvector is sufficient, because the goal of the first step is to find the eigenvector with the largest eigenvalue.

When we solve the generalization problem, we have to be cautious to following two points. First, as matrix M is symmetric and has identical paired components, the covariance should be divided by 2. Second, as the eigenvector v is composed of three vectors, a, b, and c (where $a^t a + b^t b + c^t c = 3$), we have to multiply $v^t v$ by 3.

Now we will argue that there is no difference between the constraints, in the case of two sets of variables. On this occasion, the objective is to maximize $\mathrm{Cov}(Xa, Yb)$ under the constraint $a^t a + b^t b = 2$. In a similar manner, Lagrangian function can be used to get the solution of the problem. Let us define the function L as

$$L(a, b, \lambda) = a^t X^t Y b - \lambda(a^t a + b^t b - 2). \tag{2.8}$$

By setting the partial differential of L to 0_p and 0_q,

$$\frac{\partial L}{\partial a} = X^t Y b - 2\lambda a = 0_p, \tag{2.9}$$

$$\frac{\partial L}{\partial b} = Y' X a - 2\lambda b = 0_q. \tag{2.10}$$

By solving the simultaneous equations of (2.9) and (2.10) with respect to a, b is eliminated. Consequently, we have

$$X^t Y Y^t X a = 4\lambda^2 a. \tag{2.11}$$

Here, the solution of a is an eigenvector of $p \times p$ non-negative matrix $X^t Y Y^t X$. In the same manner, the solution of b is an eigenvector 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 $X^t Y$. (2.11) is the same with the result derived under the constraint $a^t a = b^t b = 1$. Regardless of constraints, the same eigensystem is derived.

To prove that the above idea is correct, we are going to apply the same idea to the case of data matrix with two sets of variables. If that idea can be applied to the case of two data sets, we can infer that it would work for the case of more than two data sets.

To solve simultaneous equations of (2.9) and (2.10), and to obtain eigensystem, let us define matrix E_2 and D_2 as follows.

$$E_2 = \begin{bmatrix} X^t X & X^t Y \\ Y^t X & Y^t Y \end{bmatrix}, \quad D_2 = \begin{bmatrix} X^t X & \\ & Y^t Y \end{bmatrix}, \quad \lambda = \text{constant}, \quad v = \begin{bmatrix} a \\ b \end{bmatrix},$$

where $v^t v = 2$ ($\because a^t a + b^t b = 2$).

By the use of (2.9), and (2.10), we consider

$$(E_2 - D_2)v = 2\lambda v. \tag{2.12}$$

If we denote M_2 as $E_2 - D_2$, we obtain the following eigensystem.

$$M_2 v = 2\lambda v, \tag{2.13}$$

where $M_2 = \begin{bmatrix} 0 & X^t Y \\ Y^t X & 0 \end{bmatrix}$.

Thus, when we use SVD of matrix M_2 to obtain eigenvalues and eigenvectors, it means SVD of $X^t Y$ and $Y^t X$.

3. Quantification Algorithm

The quantification method for PLS correlation with many sets of variables are exactly the same with the case of two sets of variables (for the two sets of data, refer to the Huh *et al.* (2007)). Finding weight vectors is very complicated (as shown above) when compared to PLS correlation with two sets of variables.

For convenience, we will consider three sets of variables, $X(n \times p)$, $Y(n \times q)$ and $Z(n \times r)$ for quantification algorithm for generalization of PLS correlation. We assume that data are centered and scaled. We project X onto $Xa = s\,(n \times 1)$, Y onto $Yb = t\,(n \times 1)$, and Z onto $Zc = u\,(n \times 1)$. The quantification procedures are as follows (We use the following notations).

Notations

- K: data matrix with many sets of data matrix
- X, Y, Z: each set of data matrix
- M: square matrix (K^tK) without diagonal
- a, b, c: weight vectors obtained from SVD
- s, t, u: score vectors of data set X, Y
- g_X: loading vector of data set X
- g_Y: loading vector of data set Y
- g_Z: loading vector of data set Z
- $\hat{X}$: predicted value of variables in data set X
- $\hat{Y}$: predicted value of variables in data set Y
- $\hat{Z}$: predicted value of variables in data set Z
- number in subscript: PLS cycle

Cycle 1

Step 1: Find weight vectors and score vectors

Find weight vectors, a_1, b_1, and c_1 in the manner of maximizing $\mathrm{Cov}(Xa_1, Yb_1) + \mathrm{Cov}(Xa_1, Zc_1) + \mathrm{Cov}(Yb_1, Zc_1)$ under the constraints of $a_1^t a_1 + b_1^t b_1 + c_1^t c_1 = 3$. By SVD of matrix M, we obtain eigenvector of M. By multiplying eigenvector of M by square root 3, weight vectors, a_1, b_1 and c_1 can be obtained. When we obtain a_1, b_1 and c_1, we can calculate s_1, t_1 and u_1 by multiplying X_1, Y_1 and Z_1 by a_1, b_1 and c_1 respectively. Accordingly, we can obtain score vector $X_1a_1 = s_1$, $Y_1b_1 = t_1$ and $Z_1c_1 = u_1$. In this step, if eigenvalues of matrix M_1 are all zero or negative, we have to stop this quantification process. If one of the eigenvalues at least is positive, we continue the process.

Step 2: Find loading vectors

Obtain $\hat{X}_1$, $\hat{Y}_1$ and $\hat{Z}_1$ by regressing X_1 on s_1, Y_1 on t_1 and Z_1 on u_1 respectively. We can find loading vectors in this step. Loading vectors of variables on scores are the regression coefficients. They are as follows.

$$\hat{X}_1 = s_1 \left(s_1^t s_1\right)^{-1} s_1^t X_1 \ \left(= s_1 g_{1.X}^t\right), \quad \text{where } g_{1.X}^t = \left(s_1^t s_1\right)^{-1} s_1^t X_1,$$
$$\hat{Y}_1 = t_1 \left(t_1^t t_1\right)^{-1} t_1^t Y_1 \ \left(= t_1 g_{1.Y}^t\right), \quad \text{where } g_{1.Y}^t = \left(t_1^t t_1\right)^{-1} t_1^t Y_1,$$
$$\hat{Z}_1 = u_1 \left(u_1^t u_1\right)^{-1} u_1^t Z_1 \ \left(= u_1 g_{1.Z}^t\right), \quad \text{where } g_{1.Z}^t = \left(u_1^t u_1\right)^{-1} u_1^t Z_1.$$

Table 3.1. Quantification formulas of PLS correlation for columns (variables)

	Dimension 1	Dimension 2
X variables	$x_j^t s_1^*$	$x_j^t s_2^*$
Y variables	$y_k^t t_1^*$	$y_k^t t_2^*$
Z variables	$z_l^t u_1^*$	$z_l^t u_2^*$

Note: $s_1^* = s_1/||s_1||$, $s_2^* = s_2/||s_2||$, $t_1^* = t_1/||t_1||$, $t_2^* = t_2/||t_2||$, $u_1^* = u_1/||u_1||$, $u_2^* = u_2/||u_2||$

Table 3.2. Quantification formulas of PLS correlation 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	$Y_1 b_1 = t_1$	$Y_2 b_2 = t_2$
Observations in data matrix Z	$Z_1 c_1 = u_1$	$Z_2 c_2 = u_2$

Here, $g_{1.X}^t$, $g_{1.Y}^t$ and $g_{1.Z}^t$ are loading vectors for each set of variables.

Step 3: Deflate the data

Deflate X_1, Y_1 and Z_1 in the following manner.

$$X_2 = X_1 - \hat{X}_1,$$
$$Y_2 = Y_1 - \hat{Y}_1,$$
$$Z_2 = Z_1 - \hat{Z}_1.$$

Cycle 2

Step 4: Find weight vectors and score vectors

Find a_2, b_2 and c_2 in the manner of maximizing $\mathrm{Cov}(X_2 a_2, Y_2 b_2) + \mathrm{Cov}(X_2 a_2, Z_2 c_2) + \mathrm{Cov}(Y_2 b_2, Z_2 c_2)$ under the constraint of $a_2^t a_2 + b_2^t b_2 + c_2^t c_2 = 3$. SVD of matrix M_2 that is derived from simultaneous equations is needed to obtain a_2, b_2 and c_2. If we obtain a_2, b_2 and c_2, we can calculate s_2, t_2 and u_2 by multiplying X_2, Y_2 and Z_2 by a_2, b_2 and c_2, respectively, like Step 1.

Step 5: Finding loading vectors

Obtain $\hat{X}_2$, $\hat{Y}_2$ and $\hat{Z}_2$ by regressing X_2 on s_2, Y_2 on t_2 and Z_2 on u_2 respectively. We can obtain loading vectors in this step.

Through the cycle of getting suitable number of components, we can get the coordinates of the variables(loading vectors) and observations(score vectors). Thus, columns $x_j\,(j = 1, 2, \ldots, p)$ of X can be pointed on the linear space $P_j : (x_j^t s_1^*, x_j^t s_2^*, \ldots)$ generated by $s_1, s_2, \ldots$. Columns $y_k\,(k = 1, 2, \ldots, q)$ of Y can be pointed on the linear space $Q_k : (y_k^t t_1^*, y_k^t t_2^*, \ldots)$ generated by $t_1, t_2, \ldots$. Columns $z_l\,(l = 1, 2, \ldots, r)$ of Y can be pointed on the linear space $R_l : (z_l^t u_1^*, z_l^t u_2^*, \ldots)$ generated by $u_1, u_2, \ldots$. Here $s_1^* = s_1/||s_1||$, $t_1^* = t_1/||t_1||$ and $u_1^* = u_1/||u_1||$.

4. Numerical Example

4.1. Data description

The data shown as an example here are the survey results of the Chinese automobile market. We showed three sets of data that influence brand performance. They are data for the property evaluation of the automobile, the data for property evaluation of dealer service, and the data for

Table 4.1. 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 4.2. Property list of automobile brands

1. proper engine displacement/power	2. engine type (V6 or diesel engine)
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. fuel efficiency	24. maintenance efficiency
25. manufacturer impression	26. place of origin
27. availability of parts	28. cargo capacity
29. antitheft device	30. overall safety
31. environmental protection	32. sales service
33. after-sales service	34. lead time

Table 4.3. Property list of automobile dealer service

1. convenience of visiting	2. vehicle display
3. test drive	4. salesperson's knowledge about vehicle
5. courtesy/friendliness/honesty of salesperson	6. clear/accurate explanation of all documents
7. financial arrangements	8. time period for the final delivery
9. state of your vehicle at delivery	10. relationship after purchase
11. sufficient supply of advertising materials	

property evaluation of repair service on 50 automobile brands (= companies) that are collected from the survey done in 2006.

Automobile buyers of each brands evaluated the properties of car they bought, properties of dealer service they visited and properties of repair service experienced. The evaluation data are averaged based on the brands that automobile buyers experienced.

We consider the data matrix K with 53 variables and 50 observations (companies). Data matrix K consists of three sets of variables such as $X\,(50 \times 34)$, $Y\,(50 \times 11)$ and $Z\,(50 \times 8)$. Here, X is a data set for the consumer evaluation of properties for automobile repair services. Y is a data set

Table 4.4. Properties list of automobile repair services

1. the number of 4s shop for repair service is proper
2. convenience of location of 4s shop for repair service
3. fairness of charges for service work performed
4. quick repair work
5. repair equipment and facility are excellent
6. quality of work performed (skills) are excellent
7. kindness of the repair person
8. repair service center environment (customer lounge and cleanness)

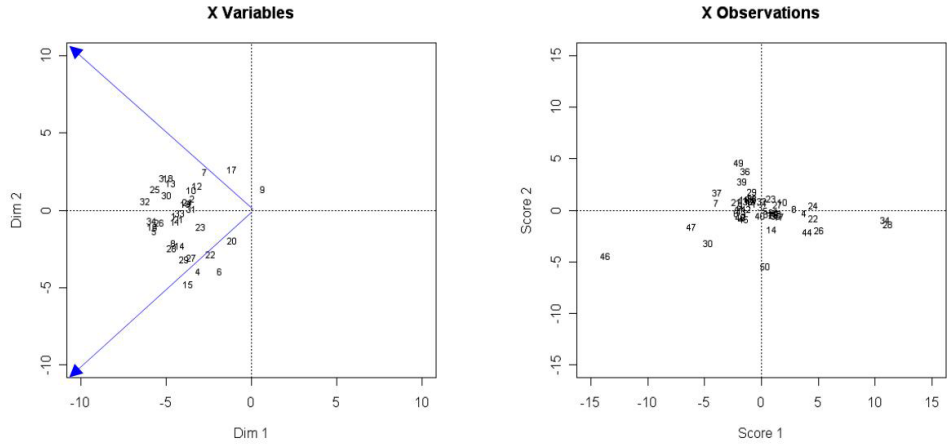

Figure 4.1. Plots of variables and observations of X by PLS generalization

for the consumer evaluation of automobile dealer service of the company, and Z is a data set for the consumer evaluation of automobile repair service of the company. Here, data sets X, Y, and Z are collected on the seven point scale (from point 1 to point 7) that is scaled and centered for the analysis. The details are listed in Table 4.1, Table 4.2, Table 4.3 and Table 4.4.

4.2. Interpretation of quantification result

The left part of Figure 4.1 is a projection of variables (property evaluation of automobile brands) in data set X onto the space generated by score vector s_1 and s_2. Each number in the figure means the variables of data set X. The right part of Figure 4.1 is a projection of observations (brands) onto the same space.

The plots in data set X and the plots in data set Y are scattered with similar direction. See Figure 4.2. We can find that data set X and data set Y have a positive relationship in the shape. The plots in data set Z are scattered with the opposite direction to the plots in data set X and the plots in data set Y. See Figure 4.3. It means that data set Z has a negative relationship with data set X and data set Y.

An interesting phenomenon is found in the right part of the Figure 4.1. The X variables are divided into two groups on the direction. The variables of the first group gather around variable 30 (overall safety). They are 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

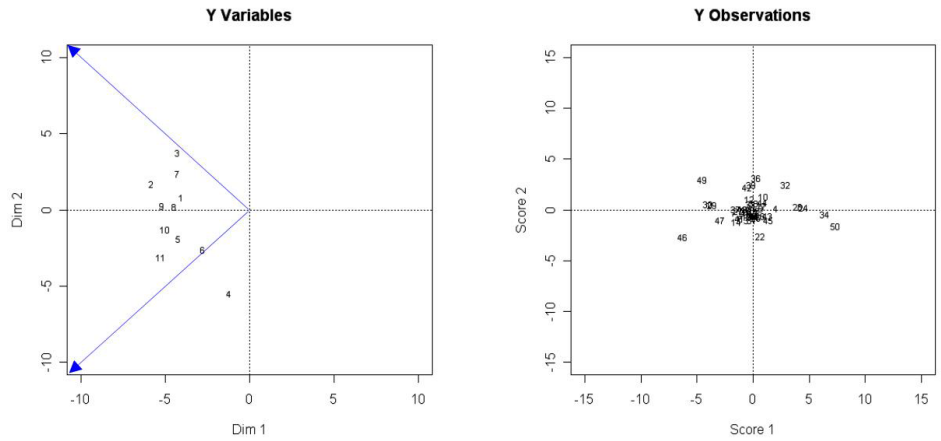

Figure 4.2. Plots of variables and observations of Y by PLS generalization

trading price). They are variable 20 (price), variable 15 (convenience to load and unload cargoes), variable 29 (antitheft device) 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 differences in the first axis among the observations (brands). That is, the differences among the brands occur only in the first axis.

We can combine the plots of X variables with X observations. Brand 46 (Beijing Futon) is very far in the direction of the second group of the X variables. It can be interpreted that brand 46 is evaluated most positively in the second group of variables. Brand 47 (Beijing Auto) and brand 30 (Jilin Tongtian) have a same direction with the second group; however, brand 37 (Huachen BMW) has a same direction with the first group of variables.

In addition, brand 34 (SGM Wuling) and brand 28 (Jiangnan Auto) are in the opposite direction to the other variables. It seems that they are poorly evaluated in the properties. Brand 44 (Jiao Auto) and brand 26 (Changan Motor) have similar position with brand 34 and brand 28.

The left part of Figure 4.2 is a projection of variables (property evaluation of automobile dealer service) in data set Y onto the space generated by score vector t_1 and t_2. Each number in the figure means the variables of data set Y. And the right part of Figure 4.2 is a projection of observations (brands) onto the same space.

The overall plots of data set Y are very similar to data set X. The Y-variables can be divided into two groups. The variables of the first group gather around variable 1 (convenience of visiting) and variable 2 (display of vehicle). The variables of the second group gather around the variable 5 (courtesy/friendliness/honesty of salesperson) and variable 11 (relationship after purchase). We can give the meaning of 'dealer's basic function' to the first group of variables, considering the meaning of the variables. We can infer the meaning of 'dealer's additional function' from the second groups of variables; however, only the variable 4 (salesperson's knowledge about vehicle) locates far and solely from the other groups of variables.

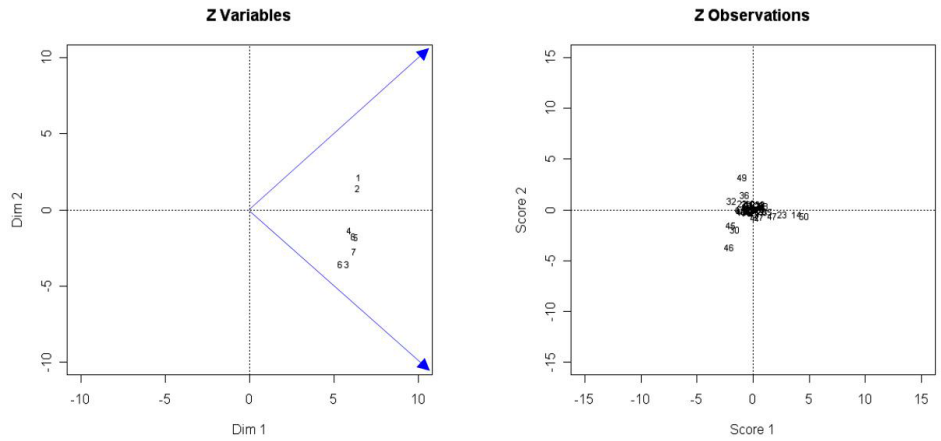

Figure 4.3. Plots of variables and observations of Z by PLS generalization

Brand 46 (Beijing Futon) has the same direction with the second group of variables and it can be interpreted that brand 46 is evaluated most positively in the properties of 'dealer's additional function'; however, brand 49 (Baolong Auto) has the same direction with the first group of variables. Similarly we can interpret that brand 49 is evaluated most positively in the properties of 'dealer's basic function'.

Brand 50 (Mercedes-Benz) has an opposite direction with a overall plots of the Y variables. It can be interpreted that brand 50 is evaluated most negatively; in addition, brand 34 (SGM Wuling) has a similar position to brand 50.

The left part of Figure 4.3 is a projection of variables (property evaluation of automobile repair service) in data set Z onto the space generated by score vector u_1 and u_2. Each number in the figure means the variables of data set Z. And the right part of Figure 4.3 is a projection of observations (brands) onto the same space. Very similarly to the plots of X and Y, the variables of Z are scattered along the second axis.

The Z-variables are also divided into two groups. The first group consists of variable 1 (the number of 4s shop) and variable 2 (convenience of location). The second group consists of other variables. We can give the meaning of 'convenience of the repair service' to the first group, and the meaning of 'quality of the repair service' to the second group.

The overall plots of data set Z are somewhat different from to data set X and Y. The observations (brands) are dense around the second axis and the first axis. In the case of properties of the repair service of the brands, automobile buyers do not perceive the brands differently. Brand 50 (Mercedes-Benz) and brand 14 (Guangzhou Toyota) have the same direction with the plots of the variables in data set Z. In contrary, brand 49 (Baolong Auto) has the opposite direction to the 'quality of the repair service'. Brand 46 (Beijing Futon) has the opposite direction to the 'convenience of the repair service'.

5. Summary and Future Research Direction

This research proposes a quantification algorithm for PLS method. We propose how to quantify

PLS methods that have several sets of variables based on singular value decomposition. To derive the form of singular value decomposition in the data with more than 2 sets more easily, we used the constraint, $a^t a + b^t b + c^t c = 3$ not $a^t a = 1$, $b^t b = 1$, and $c^t c = 1$. However, there is no difference. We showed it by the use of 2 data sets case, because it is very complicated to prove it with 3 data sets.

After getting the form of singular value decomposition, a similar process we suggested in an earlier paper (Huh *et al.*, 2007) was adopted to quantify the many data sets. Consequently, the quantification technique for a PLS method gives us a better understanding of the structure of variables and observations. The quantification technique proposed here is very useful when there are several sets of variables.

We can consider another quantification method for PLS correlation. When there are sets of variables, the number of extracted score vectors would be the number of set variables. In that case the quantification of the variables and observations on the linear spaces that are generated by score vectors will be complex. Let us consider following case. When there are M sets of variables, the number of score vectors extracted in the first PLS cycle will be M. When we perform the second cycle, the number of score vectors will be $2M$. If M is large, the representation or visualization of the data set will be very complicated.

The idea of quantification is to represent or visualize the variables and observations on the reduced space. For that reason, when there are so many sets of variables to be quantified, that kind of quantification might be meaningless. Thus, we can consider another quantification algorithm for PLS method in case there are so many sets of variables.

Wold *et al.* (1987, recited in Westerhuis *et al.*, 1998) proposed two types of so-called multi-block case algorithm for analyzing the interrelationship among the blocks (they called set of variables as block). The first one is consensus PCA (= CPCA) and the other one is hierarchical PCA(HPCA).

The difference between CPCA and HPCA lies in selecting the starting super score; however, in the case of CPCA, the resulting super score will be variant depending on how starting super score is selected. HPCA resolves this problem. As eigenvector corresponding to the largest eigen value in SVD of $X^t X$ is used as a starting super score, it is very stable. Westerhuis *et al.* (1998) showed this phenomenon with a Monte Carlo simulation. They compared two cases. The one is the case where one of the blocks has strong direction. In such a case, super block has a strong relationship with the block that has a strong direction; however, such a phenomenon did not happen when the directions spread. Even though they proposed how to manage the case in which there are multiblocks, their propositions were only algorithmic and not based on the theoretical backgrounds. For that reason it is critical to manage the case there are multiblocks based on theoretical backgrounds.

APPENDIX

Table A.1. The score vectors of X, Y and Z variables

brands	Score vectors of X		Score vectors of Y		Score vectors of Z	
	score 1	score 2	score 1	score 2	score 1	score 2
1	0.3663	0.7626	0.1382	0.2940	−0.0386	0.5458
2	−1.5049	−0.5742	−0.2576	−0.5636	−0.7304	0.0853

3	−0.0475	0.3127	−0.6010	−0.9593	0.8076	−0.1807
4	3.7787	−0.2848	2.0303	0.1928	0.4720	0.2237
5	0.3323	0.0015	0.3642	−0.0843	0.2829	0.0450
6	−2.1816	−0.2820	−1.1128	−0.7887	−0.4496	−0.0337
7	−3.9001	0.7524	−1.6992	−0.2733	−0.3462	−0.1066
8	2.9320	0.1882	0.6982	0.1934	−0.6932	0.2829
9	−1.8412	0.5180	−0.2684	0.2503	−0.1817	−0.2439
10	1.9234	0.8354	0.9680	1.3497	−0.2317	0.6050
11	−0.7746	0.6578	0.2284	−0.6854	−0.2028	−0.1294
12	−1.0193	0.9393	−0.2774	1.1170	−0.2266	0.6583
13	−1.5944	0.8734	−0.3145	−0.6073	−0.5303	0.0840
14	0.9663	−1.8595	−1.4687	−1.1255	3.9477	−0.4441
15	1.0174	−0.2982	−0.4900	−0.6907	−0.4608	−0.1471
16	1.3862	−0.3349	0.0208	−0.3007	0.3169	0.0055
17	1.6034	−0.5492	0.8091	0.3806	−0.3163	0.2421
18	−0.8669	0.9461	−0.5261	−0.3285	−0.2333	0.2086
19	−1.4766	1.0658	−0.5795	0.0276	−0.5132	0.4340
20	1.0762	−0.1999	−0.1703	−0.5840	−0.1187	0.0082
21	−2.2115	0.8274	−0.7113	−0.0740	0.0733	0.0170
22	4.5771	−0.7649	0.6452	−2.5679	−0.9690	0.6866
23	0.8801	1.1975	−0.0922	−0.2031	2.6345	−0.4476
24	4.5622	0.5207	4.5172	0.2437	−0.2055	0.1910
25	1.1036	−0.3842	0.3100	−0.7643	0.6410	0.0478
26	5.1196	−1.9178	0.6525	−0.6194	0.6151	0.4882
27	1.4711	0.5784	−1.3455	−0.0751	0.5769	−0.6795
28	11.1261	−1.3570	4.0081	0.3572	0.9204	0.3958
29	−0.8320	1.8618	−3.6641	0.4927	0.8192	−0.1000
30	−4.6295	−3.1900	−3.9764	0.5589	−1.5881	−1.9237
31	0.6215	−0.2972	−0.1072	−0.9680	−0.3520	−0.1209
32	0.0770	0.8914	2.9164	2.4799	−1.8673	0.9065
33	−0.9141	1.0553	−0.2314	0.5575	−0.2459	0.0791
34	10.8824	−0.9269	6.4360	−0.4241	0.6360	0.4529
35	1.3115	−0.5174	−0.0967	−0.5627	1.3280	−0.1548
36	−1.3781	3.8311	0.2725	3.1535	−0.7308	1.4949
37	−3.8860	1.7445	−1.5129	0.1192	−0.5286	0.0388
38	−0.8294	1.1516	0.0747	0.7010	0.6766	0.5670
39	−1.6542	2.8625	−0.1657	2.5099	−0.5889	0.5496
40	−0.0649	−0.5287	−0.0068	−0.1310	−1.0133	−0.1690
41	−1.8691	0.1519	−1.1821	−0.7897	0.2352	−0.7687
42	−1.3182	0.1399	−0.5429	2.2391	0.0480	−0.3953
43	−1.7102	−0.0751	1.2909	−0.5829	−1.1701	0.0934
44	4.0603	−2.0520	0.9010	0.7748	−0.6845	−0.2508
45	−1.5126	−0.8532	1.3843	−1.0048	−1.9323	−1.5297
46	−13.6603	−4.4101	−6.2431	−2.6579	−2.1269	−3.6249
47	−6.0768	−1.5931	−2.8835	−1.0332	1.7277	−0.6202
48	−1.7975	−0.6652	−0.9722	0.0797	−1.0814	−0.0519
49	−1.9704	4.6466	−4.5104	2.9712	−0.9774	3.2727
50	0.3472	−5.3985	7.3437	−1.5943	4.5763	−0.5874

Table A.2. The loading vectors of X variables

variables	Loading vectors of X variables	
	dimension 1	dimension 2
$X1$	−4.500	−0.356
$X2$	−3.434	0.753
$X3$	−5.216	2.075
$X4$	−3.102	−3.936
$X5$	−5.682	−1.347
$X6$	−1.855	−3.918
$X7$	−2.709	2.510
$X8$	−4.565	−2.077
$X9$	0.667	1.390
$X10$	−3.478	1.313
$X11$	−4.403	−0.701
$X12$	−3.153	1.593
$X13$	−4.727	1.794
$X14$	−4.185	−2.296
$X15$	−3.669	−4.747
$X16$	−5.782	−1.041
$X17$	−1.105	2.641
$X18$	−4.838	2.121
$X19$	−3.821	0.436
$X20$	−1.097	−1.922
$X21$	−4.234	−0.553
$X22$	−2.398	−2.810
$X23$	−2.961	−1.044
$X24$	−3.731	0.538
$X25$	−5.644	1.379
$X26$	−5.366	−0.750
$X27$	−3.497	−3.036
$X28$	−4.655	−2.421
$X29$	−3.931	−3.130
$X30$	−4.953	1.001
$X31$	−3.495	0.127
$X32$	−6.220	0.595
$X33$	−4.165	−0.169
$X34$	−5.795	−0.653

Table A.3. The loading vectors of Y variables

variables	Loading vectors of Y variables	
	dimension 1	dimension 2
$Y1$	−4.031	0.806
$Y2$	−5.798	1.714
$Y3$	−4.287	3.754
$Y4$	−1.191	−5.462
$Y5$	−4.231	−1.886
$Y6$	−2.769	−2.584
$Y7$	−4.258	2.400
$Y8$	−4.482	0.209
$Y9$	−5.196	0.291
$Y10$	−5.001	−1.297
$Y11$	−5.241	−3.122

Table A.4. The loading vectors of Z variables

variables	Loading vectors of Z variables	
	dimension 1	dimension 2
Z1	6.505	2.183
Z2	6.413	1.458
Z3	5.763	−3.565
Z4	5.939	−1.301
Z5	6.301	−1.789
Z6	5.396	−3.529
Z7	6.210	−2.709
Z8	6.155	−1.699

References

Han, S.-T. (1995). Quantification Approach to Ranked Data Analysis, Doctoral Dissertation, Korea University.

Helland, I. (2005). Partial least squares regression, *The Encyclopedia of Statistical Sciences*, Second Edition (edited by Kotz), 5957–5962.

Huh, M.-H. (1999). *Quantification Methods for Multivariate Data*, Free Academy, Seoul.

Huh, M.-H., Lee, Y. and Yi, S. (2007). Visualizing (X, Y) data by partial least squares method, *Korean Journal of Applied Statistics*, **20**, 345–355.

Husson, F. and Pages, J. (2005). Scatter plot and additional variables, *Journal of Applied Statistics*, **32**, 341–349.

Park, M. and Huh, M.-H. (1996). Quantification plots for several sets of variables, *The Journal of Korean Statistical Society*, **25**, 589–601.

Westerhuis, J. A., Kourti, T. and Macgregor, J. F. (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models, *Journal of Chemometrics*, **12**, 301–321.

Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, **2**, 37–52.

Yi, S. K. (2007). *Quantification Method for Partial Least Squares and its Generalization*, Doctoral Dissertation, Korea University.