응용통계연구 제 **20**권 2호, *2007*년, *pp. 345–355*

PLS 기법에 의한 (X,Y) 자료의 시각화

허명회[1] 이용구[2] 이성근[3]

요 약

PLS 회귀는 q-변량의 Y 변수에 대한 회귀에서 p-변량의 X 변수가 다중공선성의 문제를 갖는 경우에도 적용가능한 방법이다. 특히 X 변수의 수 p가 관측개체 수 n보다 큰 경우에 적용 가능하여 계량화학(chemometrics) 분야에서 근적외선 분광기(near-infrared spectroscopy) 자료에 대한 표준적 분석 방법으로 활용되고 있다. 이 연구에서 우리는 PLS 회귀의 방법론을 정리하고 이를 활용한 p개의 X 변수들과 q개의 Y 변수들의 동시 시각화를 위한 두 가지의 수량화 방법을 제안한다.

주요용어: PLS 회귀, 자료 시각화, 수량화 플롯.

1. 들어가며

전통적인 선형회귀에서는 Y를 $X_1, X_2, \ldots, X_p$로 표현한 선형모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \quad \varepsilon \sim (0, \sigma^2)$$

을 적합할 필요가 있는데 이를 위하여 p보다 큰 n개의 관측이 있어야 한다. 실제 안정된 모형을 얻기 위해서는 n이 p보다 훨씬 커야 한다. 따라서 n이 p보다 크지 않거나 그렇지 않더라도 n이 p에 비해 그다지 크지 않으면 최소제곱회귀는 불가능하거나 불안정하게 되는 문제가 있다.

PLS(Partial Least Squares) 회귀는 이와 같은 경우에서도 기능할 수 있는 선형회귀적 방법이다. 이 방법은 1966년 Herman Wold에 의하여 계량경제적(econometric) 기법으로 창안되었는데 그 이후 여러 분석화학자들이 받아들여 계량화학적(chemometric) 기법으로 활용하고 있다. PLS 회귀에 대한 최근 20여년에 걸친 연구는 수백 편에 이르고 있어 간단히 정리하기는 어렵다. 1–2편을 고른다면 Helland(2006)의 Encyclopedia of Statistical Sciences 소개 논문과 Rosipal과 Kramer(2006)의 개괄논문을 뽑을 수 있다. 국내 이론 연구자로는 2001년이후 여러 연구결과를 낸 김종덕 교수가 있고 (김종덕 2004; Kim 2001, 2003a, 2003b, 2003c) 응용 연구로는 박성현 등(1999), 전치혁 등(2006) 등이 있다.

1) (136-701) 서울특별시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수
E-mail: stat420@korea.ac.kr
2) (156-756) 서울특별시 동작구 흑석동, 중앙대학교 통계학과, 교수
E-mail: leeyg@cau.ac.kr
3) (136-742) 서울특별시 경북구 동선동, 성신여자대학교 경영학과, 부교수
E-mail: yisk@sungshin.ac.kr

우리는 이 연구에서 PLS 회귀의 방법론을 정리하고 이를 활용하여 p개의 X 변수들과 q개의 Y 변수들의 연관성을 동시 시각화하는 자료 탐색적 방법을 제안하고자 한다. 따라서 제안방법론은 정준상관분석(canonical correlation analysis)을 활용한 자료 시각화(Park과 Huh, 1996a)와 동일한 목적으로 활용가능하다. 다만 정준상관분석은 p와 q에 비교하여 n이 큰 경우에만 적용 가능하지만 이 연구에서 제안하는 시각화 방법론에는 그런 제약이 붙지 않는다는 점이 다르다. 정준상관분석과 PLS의 관계에 대하여는 Wegelin(2000), Barker와 Rayens(2003), Rosipal과 Kramer(2006) 등이 검토한 바 있다.

2. PLS 회귀 방법론

p개의 설명변수와 q개의 반응변수, 이에 대한 n개의 관측이 있다고 하자. 이를 2개의 행렬 $X(n \times p)$와 $Y(n \times q)$로 표기하자. 특별한 언급이 없는 한, X와 Y가 열 별로 중심화되고 척도화되어 있음을 가정하지만 반드시 요구되는 가정은 아니다.

PLS 회귀 방법론은 최대 공분산을 갖는 p개 설명변수의 선형결합과 q개 반응변수의 선형결합을 찾는 데에서 시작한다. 즉,

$$\text{maximize (wrt } b \text{ and } c) \ \mathrm{Cov}(Xb, Yc), \quad \text{여기서} \ \ b : p \times 1, \ c : q \times 1 \tag{2.1}$$

이다. 단, (2.1)의 공분산이 b와 c의 방향(direction) 뿐 아니라 크기(norm)에도 의존하므로 제약조건으로

$$b^t b = 1, \quad c^t c = 1 \tag{2.2}$$

을 고려한다. 따라서 최적화 함수 $h()$를

$$h(b, c, \lambda_1, \lambda_2) = b^t X^t Y c - \lambda_1 (b^t b - 1) - \lambda_2 (c^t c - 1)$$

로 정의하고 이를 각각 b와 c로 편미분하여 0_p과 0_q로 놓음으로써

$$X^t Y c - 2\lambda_1 b = 0_p, \quad Y^t X b - 2\lambda_2 c = 0_q$$

를 얻는다. c를 소거하면

$$X^t Y Y^t X b = 4\lambda_1 \lambda_2 b$$

가 되므로 $p \times p$ 비음 정부호 행렬 $X^t Y Y^t X$의 고유벡터가 b의 해가 된다. 마찬가지로 $q \times q$ 비음 정부호 행렬 $Y^t X X^t Y$의 고유벡터가 c의 해가 된다. 그러므로 b와 c는 $p \times q$ 행렬 $X^t Y$의 비정칙값 분해(SVD, singular value decomposition)로부터 얻어질 수 있다. 왜냐하면

$$X^t Y = U D_\mu V^t; \quad U = (u_1, u_2, \ldots), \quad V = (v_1, v_2, \ldots), \quad \mu_1 \geq \mu_2 \geq \cdots$$

로부터 (여기서 $u_1, u_2, \ldots$는 r개의 직교정규 $p \times 1$ 열 벡터이고 $v_1, v_2, \ldots$는 r개의 직교정규 $q \times 1$ 열 벡터, r은 행렬 $X^t Y$의 계수(rank)임),

$$b = u_1, \quad c = v_1$$

이 얻어지기 때문이다. PLS 회귀에서는 b와 c를 가중치 벡터(weight vector)라고 한다.

Y의 차원 크기가 $n \times 1$인 경우, 즉 $Y = y$인 경우 어렵지 않게

$$b = X^t y / \parallel X^t y \parallel$$

임을 보일 수 있다. 따라서 반응변수가 1개인 경우엔 비정칙값 분해를 하지 않아도 된다.

이제 Xb를 $s(n \times 1)$로 표기하고 점수 벡터(score vector)라고 부른다. Y를 s에 회귀함으로써 Y적합

$$\widehat{Y} = s(s^t s)^{-1} s^t Y = s g_Y^t, \quad g_Y^t = (s^t s)^{-1} s^t Y \tag{2.3}$$

를 얻는다. $g_Y(q \times 1)$를 Y-부하(loading) 벡터라고 부른다.

형식적으로 본다면 PLS 회귀는 s를 설명변수로 하는 Y에 대한 선형회귀이다(Rosipal과 Kramer, 2006). 그런데 $s = Xb \ (= x_1 b_1 + \cdots + x_p b_p, \ x_j : n \times 1)$에서 계수 벡터 b의 결정에 X뿐 아니라 Y가 고려되므로 $\widehat{Y}$는 Y의 선형변환으로 표현되지 않는다. 즉,

$$\widehat{Y} = AY, \quad \text{여기서} \quad A = A(Y) \text{ or } s(s^t s)^{-1} s^t \tag{2.4}$$

이다. 따라서 선형회귀에서와는 달리 $\widehat{Y}$의 분포적 행태가 쉽게 구해지지 않는다.

이상은 PLS 회귀의 1 단계에 해당한다. (2.4)를 보면 Y의 변환행렬 A는 계수(階數, rank) 1을 갖는다. 보다 정밀한 적합을 위해 PLS(partial least squares, 편최소제곱) 회귀에서는 다음 알고리즘으로 A의 계수를 증가시킨다. 편의상, 1 단계를 추가하는 알고리즘을 기술하기로 한다. 이하, $s \to s_1$, $g_Y \to g_{1,Y}$, $X \to X_1$, $Y \to Y_1$, $\widehat{Y} \to \widehat{Y}_1$로 표기한다.

1) X_1와 Y_1을 업데이트한다:

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

여기서 X_1의 적합 $\widehat{X}_1$은 s_1에 내린 X_1의 사영으로 $\widehat{X}_1 = s_1(s_1^t s_1)^{-1} s_1^t X_1$이다. $g_1 = X_1^t s_1 (s_1^t s_1)^{-1}$로 표기하면 $\widehat{X}_1 = s_1 g_1^t$가 된다. $g_1^*(p \times 1) = X_1^t s_1 / \parallel s_1 \parallel$을 X-부하(loadings)라고 한다.

2) $\mathrm{Cov}(X_2 b_2, Y_2 c_2)$를 최대화하는 b_2와 c_2를 찾는다($b_2{}^t b_2 = 1 \ \& \ c_2{}^t c_2 = 1$).

3) 새 점수벡터와 X_2 적합과 Y_2 적합을 계산한다:

$$s_2 = X_2 b_2, \quad \widehat{Y}_2 = s_2 (s_2{}^t s_2)^{-1} s_2{}^t Y_2, \quad \widehat{X}_2 = s_2 (s_2{}^t s_2)^{-1} s_2{}^t X_2.$$

4) 결과적으로

$$\widehat{Y} = \widehat{Y}_1 + \widehat{Y}_2 = s_1(s_1^t s_1)^{-1} s_1^t Y_1 + s_2(s_2^t s_2)^{-1} s_2^t Y_2 = s_1 g_{1,Y}^t + s_2 g_{2,Y}^t,$$
$$\widehat{X} = \widehat{X}_1 + \widehat{X}_2 = s_1(s_1^t s_1)^{-1} s_1^t X_1 + s_2(s_2^t s_2)^{-1} s_2^t X_2 = s_1 g_1{}^t + s_2 g_2^t$$

로 표현된다. 여기서

$$g_{2,Y}^t = (s_2^t s_2)^{-1} s_2^t Y_2, \quad g_2^t = (s_2^t s_2)^{-1} s_2^t X_2$$

이다. 이를 활용하여 $x^*(p\times 1)$를 갖는 개체의 미래값 $y^*(q\times 1)$에 대한 예측이 가능하다. 이와 같은 단계를 거듭하면 $\widehat{X}$과 $\widehat{Y}$은 r개의 성분을 갖게 된다 (r=1,2,3,...). 성분의 수를 결정하기 위하여 교차검토기법(cross-validation technique)을 활용한다.

3. 다변량 자료의 시각화 I

이 절에서는 p개의 설명변수와 q개의 반응변수로 구성된 다변량 자료 (X,Y)를 고려하자. 즉 Y를 X에 회귀하는 경우를 다룬다. p개의 변수와 q개의 변수로 상호 연관된 경우에 대하여는 다음 절에서 다룰 것이다.

2절에서 정리한 PLS 회귀에 의하면 $n\times 1$ 점수벡터 s_1, s_2 등은 상호직교하면서 X와 Y의 열이 투사되는 사영 공간의 기저를 생성한다. 이에 따라, X의 열 $x_j(j=1,2,\ldots,p)$를 s_1, s_2...에 의해 생성된 선형공간의

$$P_j \ : \ (x_j^t s_1^*, \ x_j^t s_2^*, \ \ldots) \tag{3.1}$$

에 타점할 수 있고 Y의 열 $y_k(k=1,2,\ldots,q)$는

$$Q_k \ : \ (y_k^t s_1^*, \ y_k^t s_2^*, \ \ldots)$$

에 타점할 수 있다. 여기서 $s_1^* = s_1/\parallel s_1 \parallel$, $s_2^* = s_2/\parallel s_2 \parallel$, ...이다.

(3.1)로부터 X의 열들이 사영되는 수량화점의 제1 성분은 $X^t s_1/\parallel s_1 \parallel$의 행으로부터 나옴을 알 수 있다. 이것은 (제1) X-부하와 같다. 제2 성분은 $X^t s_2/\parallel s_2 \parallel$인데

$$X(=\ X_1) = \widehat{X}_1 + X_2, \quad \widehat{X}_1 = s_1 g_1^t, \quad s_1^t s_2 = 0$$

이므로 $X^t s_2/\parallel s_2 \parallel = \ X_2^t s_2/\parallel s_2 \parallel$와 일치한다. 이것은 제2 X-부하와 일치한다. 따라서 X의 열 수량화점은 계량화학적 응용 연구에서 쓰이는 부하 플롯(loading plot)과 사실상 일치한다. 그러나 부하 플롯에 대한 새로운 해석을 제공한다는 데 의의를 찾을 수 있다.

수치적 사례 1: 근적외선(Near-Infrared, NIR) 분광기(spectroscopy) 자료에 적용한 수치예를 제시하도록 한다(출처: R의 pls library, http://www.r-project.org). 이 NIR 자료는 밀도(density; Y)가 다른 28개(=n) 폴리에틸렌 수지 絲(PET yarns)를 근적외선 분광기로 268개(=p) 스펙트럼별 빛의 흡수도를 측정한 분석화학적 실험에서 나온 것이다(q=1). 268개의 X 변수들이 흡수도라는 공통된 의미 척도를 가지므로 통계적 척도화를 하지 않기로 한다.

그림 3.1은 2개의 점수벡터 s_1,s_2에 의해 생성되는 선형공간에 사영된 268개 X 변수들과 1개의 Y 변수의 위치를 보여준다. 흥미롭게도, X 변수들이 연결된 괘적을 따라 위치하고 있고 Y 변수는 제 1축 방향과 거의 일치하는 것으로 나타나고 있다. 이에 따라 X-변수의 사영도에서 제1축 방향으로 좌우 양 끝에 있는 변수들을 추적해 보았더니, 좌측 끝에는 $X_{16},\ldots,X_{21}$이 위치하고 있고 우측 끝에는 $X_{80},\ldots,X_{88}$이 위치하고 있었다. 이는 이들 변수에 해당하는 파장대에 대한 흡수도로부터 시료의 Y(=밀도)를 알 수 있음을 의미한

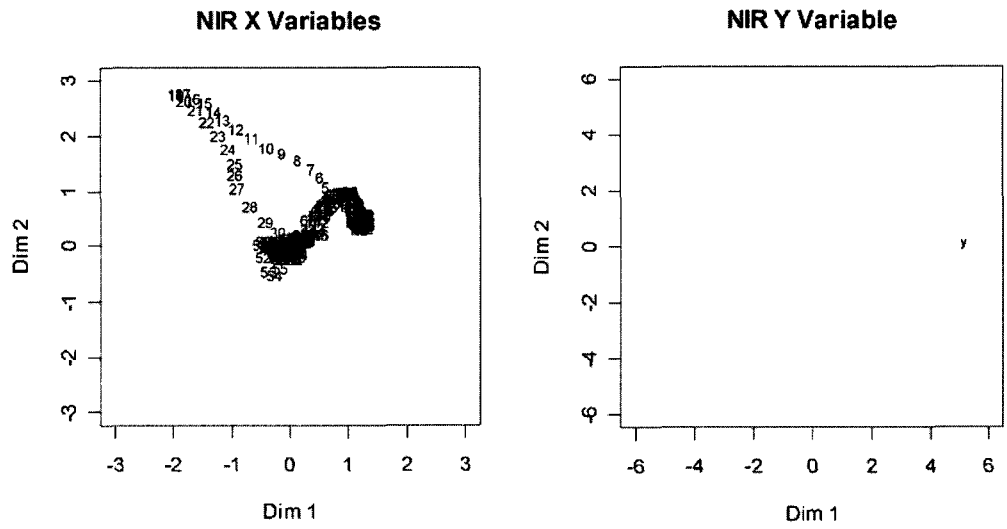

그림 3.1: NIR 자료에서 X 변수들과 Y 변수의 2차원 사영(플롯내 숫자는 X-변수 번호)

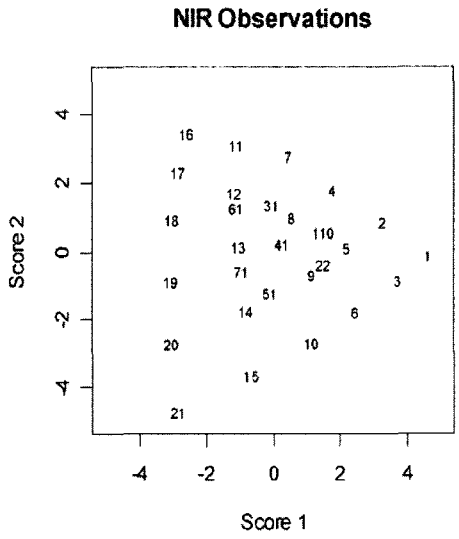

그림 3.2: NIR 자료에서 개체 점수 2차원 플롯(플롯내 숫자는 개체 번호)

다. 그림 3.2는 개체들의 2차원 X-점수플롯으로 28개 시료(=개체)의 각 차원별 점수를 볼 수 있다.

그림 3.1과 그림 3.2를 대응시킴으로써 개체들의 X-특성들을 알 수 있는데, 예를 들면

개체 1, 2, 3은 y-방향과 일치하므로 밀도가 높은 시료이고 반면 개체 16, 17, 18, 19, 20, 21은 밀도가 낮은 시료이다. 개체 16, 17, 18 등은 $X_{16}, \ldots, X_{19}$가 높고 개체 1, 2, 3 등은 큰 $X_{80}, \ldots, X_{88}$ 값을 취한다.

4. 다변량 자료의 시각화 II

이 절에서는 PLS Mode A 알고리즘을 활용하여(Rosipal과 Kramer, 2006) X와 Y가 연결된 $n \times (p+q)$다변량 자료에 적용가능한 시각화 기법을 제안하기로 한다. X와 Y는 동등한 상관관계에 있음을 가정한다. PLS 회귀에서와 마찬가지로 X의 p개 변수의 선형결합과 Y의 q개 변수의 선형결합간 공분산을 최대화하여 보자. 즉, 목적식은

$$\text{maximize (wrt } b \text{ and } c) \quad \text{Cov}(Xb, Yc), \quad \text{여기서} \quad b : p \times 1,\ c : q \times 1 \tag{4.1}$$

로 (2.1)과 같고 b와 c에 대한 제약조건으로 (2.2)를 고려한다. 이 정식화는 앞 절에서와 완전하게 같으므로 b와 c는 $p \times q$ 행렬 X^tY의 비정칙값 분해(SVD)로부터 나온다. 즉

$$b = u_1 (= b_1), \quad c = v_1 (= c_1)$$

가 된다. 여기서

$$X^tY = UD_\mu V^t; \quad U = (u_1, u_2, \ldots), \quad V = (v_1, v_2, \ldots), \quad \mu_1 \geq \mu_2 \geq \cdots$$

이다. 그러나 사영의 기저는 달라야 한다. 앞 절에서는 $s_1 = Xb_1$에 X의 모든 열과 Y의 모든 열을 사영시켰지만 X와 Y를 동등하게 놓은 이 절의 맥락에서는 X의 p개 열은 s_1에 사영시키고 Y의 q개 열은

$$t_1 = Yc_1$$

에 사영시키는 것이 자연스럽다. 이에 따라 후속 단계에서 $X(= X_1)$와 $Y(= Y_1)$는 다음과 같이 편회귀 보정된다.

1) X_1와 Y_1에서 각각의 적합($\widehat{X}_1$과 $\widehat{Y}_1$)를 감하여 업데이트한다:

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

여기서 $\widehat{X}_1$은 s_1에 내린 X_1의 사영, 즉 $\widehat{X}_1 = s_1(s_1^t s_1)^{-1} s_1^t X_1$ 이다. 그리고 $\widehat{Y}_1$은 t_1에 내린 Y_1의 사영, 즉 $\widehat{Y}_1 = t_1(t_1^t t_1)^{-1} t_1^t Y_1$이다. 따라서 X_2의 각 열은 s_1과 직교하고, Y_2의 각 열은 t_1과 직교한다.

2) $\text{Cov}(X_2 b_2, Y_2 c_2)$를 최대화하는 b_2와 c_2를 찾는다(제약: $b_2^t b_2 = 1$ & $c_2^t c_2 = 1$).

3) $$s_2 = X_2 b_2, \ \ t_2 = Y_2 c_2; \ \ \widehat{X}_2 = s_2(s_2^t s_2)^{-1} s_2^t X_2, \ \ \widehat{Y}_2 = t_2(t_2^t t_2)^{-1} t_2^t Y_2$$

이다. X_2의 각 열이 s_1과 직교하므로 s_1과 s_2는 직교한다. 마찬가지로 t_1과 t_2가 직교한다.

4) 결과적으로

$$\widehat{X} = \widehat{X}_1 + \widehat{X}_2 = s_1(s_1^t s_1)^{-1} s_1^t X + s_2(s_2^t s_2)^{-1} s_2^t X$$
$$\widehat{Y} = \widehat{Y}_1 + \widehat{Y}_2 = t_1(t_1^t t_1)^{-1} t_1^t Y + t_2(t_2^t t_2)^{-1} t_2^t Y$$

가 된다.

이에 따라, X의 열 $x_j (j = 1, 2, \ldots, p)$를 $s_1, s_2, \ldots$에 의해 생성되는 선형공간의

$$P_j : (x_j^t s_1^*,\ x_j^t s_2^*,\ \ldots)$$

에 타점할 수 있고 Y의 열 $y_k = (k = 1, 2, \ldots, q)$는

$$Q_k : (y_k^t t_1^*,\ y_k^t t_2^*,\ \ldots)$$

에 타점할 수 있다. 여기서 $s_1^* = s_1 / \parallel s_1 \parallel$, $s_2^* = s_2 / \parallel s_2 \parallel$, ..., $t_1^* = t_1 / \parallel t_1 \parallel$, $t_2^* = t_2 / \parallel t_2 \parallel$, ... 이다.

수치적 사례 2: Sensory Data는 연구자가 16개(=n) 올리브 오일 표본 각각에서 5개(=p) 물리화학적 특성을 측정하고 패널 검사원이 6개 관능 속성(=q)을 평가한 자료이다(출처: R의 pls library). 16개의 표본 중 첫 5개는 그리스 産(G1–G5), 다음 5개는 이탈리아 産(I1–I5), 그 다음 6개는 스페인 産(S1–S6)이다. PLS 회귀 적합에 앞서 모든 변수에 대하여 표준화 변환을 하였다.

그림 4.1의 왼쪽 플롯은 X의 2차원 점수벡터 s_1, s_2에 의해 생성된 선형공간에 5개 X 변수를 사영한 것이고 오른 쪽 플롯은 Y의 2차원 점수벡터 t_1, t_2에 의해 생성된 선형공간에 6개의 Y 변수를 사영한 것이다. X 변수들(Acidity, DK, K270, K232, Peroxide)이 90° 정도의 퍼진 상태로 放射되고 있는데 이것은 대체로 이들 변수들이 양의 상관을 보이고 있음을 의미한다. 그러나 Y 변수들은 2개 그룹으로 나뉜다, 즉 glossy, transp, yellow가 한 그룹을 이루고 green, syrup, brown이 다른 한 그룹을 이룬다. 두 그룹간에는 음의 상관이 있을 것으로 예상할 수 있다. K232와 Peroxide와 같은 X 변수들은 syrup과 brown과 같은 Y 변수들과 대응적 관계에, glossy와 transp와 같은 Y 변수들과는 배반적 관계에 있다. 다음 상관행렬에서 확인할 수 있다.

	Acidity	Peroxide	K232	K270	DK
yellow	−0.49	−0.41	−0.55	−0.69	−0.33
green	0.51	0.34	0.47	0.64	0.36
brown	−0.20	0.78	0.74	0.55	0.08
glossy	−0.23	−0.67	−0.70	−0.53	−0.48
transp	−0.31	−0.60	−0.62	−0.52	−0.46
syrup	0.14	0.76	0.69	0.48	0.33

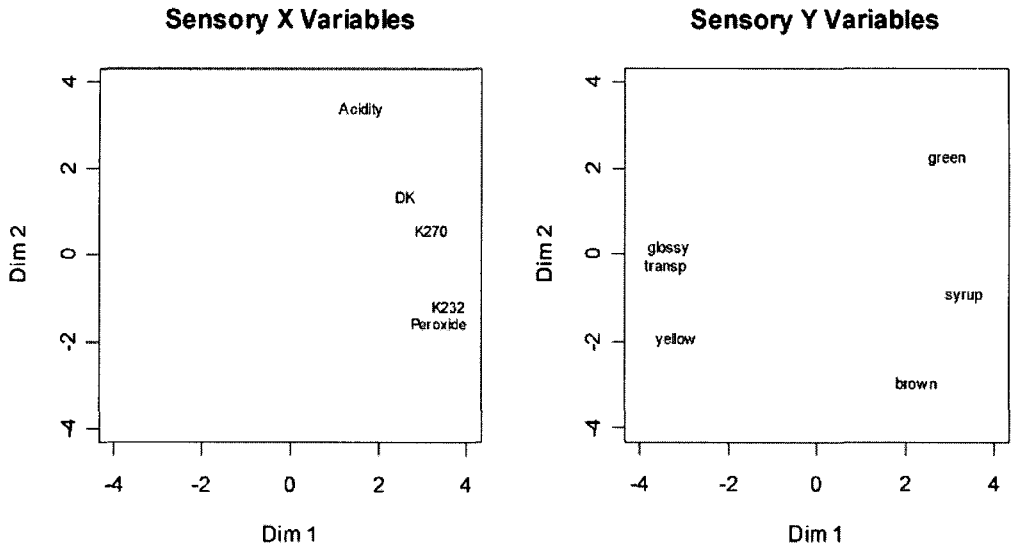

그림 4.1: sensory 자료에서 X 변수와 Y 변수의 2차원 사영

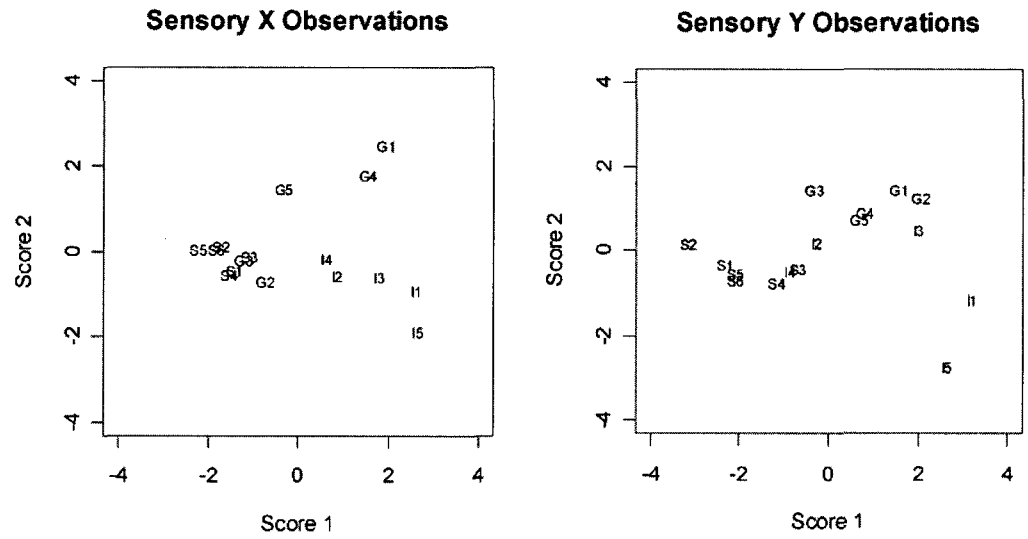

그림 4.2: sensory 자료에서 관측개체 2차원 플롯

그림 4.2는 16개 개체 표본을 2차원 점수 공간에 타점한 것이다. 산지에 따른 구분이 명확하게 나타나는데 (左=스페인 S, 右上=그리스 G, 右下=이탈리아 I) 앞 절에서와 같이 그림 4.1에 대응시켜 의미를 부여할 수 있다. 예컨대 스페인산 올리브 오일은 그리스산이나

이탈리아 산에 비해 상대적으로 모든 X 변수들(Acidity, DK, K270, K232, Peroxide)이 다소 작고 Y 특성 glossy, transp, yellow에서는 다소 큰 경향이 있다.

5. 맺음 말

우리는 이 연구에서 PLS의 두 알고리즘을 정리하고 이를 활용한 p개의 X 변수들과 q개의 Y 변수들의 동시 시각화를 위한 두 가지 수량화 방법을 제안하였다. 그 중 4절의 수량화 방법은 정준상관분석(canonical correlation analysis)과 같은 영역의 문제를 다루는 것으로 볼 수 있다(Park과 Huh 1996a, 1996b; 허명회, 1999; 최용석, 2006).

정준상관분석은 X의 p개 변수의 선형결합과 Y의 q개 변수의 선형결합간 상관을 최대화하는 방식을 취한다. 즉, 목적식이

$$\text{maximize (wrt } b \text{ and } c) \quad \text{Cov}(Xb, Yc), \quad \text{여기서} \quad b : p \times 1,\ c : q \times 1$$

이므로 목적식은 (4.1)이지만 b와 c에 대한 제약조건으로

$$b^t X^t X b = 1,\ \ c^t Y^t Y c = 1$$

를 고려하는 셈이다. 이와 비교하여 4절의 PLS-방식 수량화는 제약조건

$$b^t b = 1,\ \ c^t c = 1$$

하에서 $\text{Cov}(Xb,\ Yc)$을 최대화하는 것이므로 제약식에 있어 차이가 있다. 그런데 공분산은

$$\text{Cov}(Xb,\ Yc) = \text{Var}^{0.5}(Xb)\text{Var}^{0.5}(Yc)\text{Corr}(Xb,\ Yc)$$

이므로 PLS-방식 수량화가 하는 것은 개념적으로 1) $\text{Var}Xb$를 크게 하고 (s.t. $b^t b = 1$), 2) $\text{Var}(Yc)$를 크게 하며 (s.t. $c^t c = 1$), 3) $\text{Corr}(Xb, Yc)$를 크게 하는 것이다(Barker와 Rayens, 2003; Rosipal과 Kramer, 2006). 앞의 1)과 2)는 각각 자료행렬 X와 Y에 대한 주성분분석의 목표와 일치하므로, 4절의 PLS-방식 수량화를 다변량 자료 쌍 (X, Y)에 대한 주성분분석과 정준상관분석의 조화(harmony)로 볼 수 있다. 정준상관분석에서는 관측 수 n이 $\max(p, q)$보다 큰 경우에만 적용가능하지만 PLS-방식 수량화는 그렇지 않다는 점, 그리고 다변량 자료 쌍 (X, Y)에 대한 통계적 정보를 통합된 주성분분석과 상관분석으로부터 산출한다는 점을 PLS-방식 수량화의 특별한 장점으로 볼 수 있다.

참고문헌

김종덕 (2004). 고유벡터 기저를 이용한 회귀방법의 비교, <한국자료분석학회지>, **6**, 205–218.

박성현, 최엄문, 박창순 (1999). 편최소제곱 반응표면함수를 이용한 공정 최적화에 관한 연구, <품질경영학회지>, **27**, 237–250.

전치혁, 이혜선, 이대원, 장창환 (2006). X-선 회절 데이터에 PLS 기법을 이용한 철광석의 환원율 예측, <한국통계학회 2006년 춘계학술발표회 논문집>, **30**.

최용석 (2006). <행렬도 분석>, 부산대학교 기초과학연구원.

허명회 (1999). <다변량 수량화>, 자유아카데미.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination, *Journal of Chemometrics*, **17**, 166–173.

Helland, I. (2006). Partial least squares regression, *The Encyclopedia of Statistical Sciences*, 2nd ed., (edited by Kotz), 5957–5962.

Kim, J. D. (2001). A general weighting scheme of partial least squares regression, *Journal of the Korean Data Analysis Society*, **3**, 11–21.

Kim, J. D. (2003a). Alternative expressions of regression vector for principal component regression and partial least squares regression, *Journal of the Korean Data Analysis Society*, **5**, 17–26.

Kim, J. D. (2003b). Projection matrices for partial least squares regression and principal component regression, *Journal of the Korean Data Analysis Society*, **5**, 787–800.

Kim, J. D. (2003c). Unified non-iterative algorithm for principal component regression, partial least squares and ordinary least squares, *Journal of Korean Data and Information Science Society*, **14**, 355–366.

Park, M. and Huh, M. H. (1996a). Canonical correlation biplot, *The Korean Communications in Statistics*, **3**, 11–19.

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

Rosipal, R. and Kramer, N. (2006). Overview and recent advances in partial least squares, *Lecture Notes in Computer Science*, **3940**, 34-51, Springer-Verlag.

Wegelin, J. A. (2000). A survey of partial least squares (PLS) methods, with emphasis on the two-block case, *Technical report*, **371**, The University of Washington, Department of Statistics.

[2007년 2월 접수, 2007년 4월 채택]

Visualizing (X, Y) Data by Partial Least Squares Method

Myung-Hoe Huh[1] Yonggoo Lee[2] SeongKeun Yi[3]

ABSTRACT

PLS methods are suited for regressing q-variate Y variables on p-variate X variables even in the presence of multicollinearity problem among X variables. Consequently, they are useful for analyzing datasets with smaller number of observations compared to the number of variables, such as NIR(near-infrared) spectroscopy data in chemometrics. In this study, we propose two visualizing methods of p-variate X variables and q-variate Y variable that can be used in connection with PLS analysis.

Keywords: PLS regression, data visualization, quantification plot.

1) Professor, Department of Statistics, Korea University. Anam-dong 5-1, Seoul 136-701, Korea
E-mail: stat420@korea.ac.kr
2) Professor, Department of Statistics, Chung-Ang University. Huksuk-dong 221, Seoul 156-756, Korea
E-mail: leeyg@cau.ac.kr
3) Associate Professor, Department of Business Administration, Sungshin Women's University, Dongseon-dong 3, Seoul 136-742, Korea
E-mail: yisk@sungshin.ac.kr