한국통계학회논문집 제3권 1호
1996년 4월 pp. 1-9

다반응값 자료에 대한 biplot의 활용에 관한 연구

장 대 홍[1)]

요 약

반응표면분석에서 다반응값의 최적화 문제는 단반응값 최적화 문제보다 복잡하다. 이런 다반응값 문제에서 반응변수들이나 설명변수 상호간의 관계나 중요성 등을 평가하는 것은 중요하다. 이러한 평가를 위하여 biplot가 유용한 그림도구로 쓰일 수 있다.

1. 서론

반응표면분석(response surface methodology)에서 다반응값 문제(multiresponse problem)는 다음과 같은 범주로 나뉜다고 볼 수 있다.

1. 다반응값 모형에서의 모수추정
2. 다반응값 실험계획
3. 다반응값 모형의 적합결여검정
4. 다반응값의 최적화

이 중 다반응값의 최적화 문제는 단반응값(single response)의 최적화 문제보다 복잡하다. 보통의 경우 어느 반응값의 최적화가 다른 반응값의 최적화를 방해하거나 실행불가능하게 하기 때문이다. parameter design에서도 다특성값 문제는 단특성값 문제보다 양상이 복잡하다. 이런 다반응값 문제에서 반응변수들이나 설명변수들 상호간의 관계나 중요성 등을 평가하는 것은 매우 중요하다.

이러한 평가를 위하여 biplot가 유용한 그림도구로 쓰일 수 있다. biplot는 자료행렬을 비정칙값 분해(SVD(singular value decomposition))를 이용하여 행표시자(row marker)와 열표시자(column marker)로 나타내어 저차원(2차 내지 3차) 행렬로 근사시켜 2차원 평면이나 3차원 공간상에 그림으로 나타내어 자료행렬상의 각 변수와 관측값들 상호간의 관련성을 한 눈에 알아볼 수 있게 하는 그림도구이다. biplot는 Gabriel(1971)이 제안한 이후로 최근까지도 여러 학자에 의하여 연구되고 있다.(Corsten과 Gabriel(1976), Gabriel(1978), Gabriel과 Zamir(1979), Gabriel(1981), Gabriel과 Odoroff(1985), Gower와 Harding(1988), Gower(1990), Daigle과 Rivest(1992), Greenacre(1993)). Smith와 Cornell(1993)은 다반응값 혼합물 실험자료에 이 biplot를 이용하여 반응변수들이나 설명변수들 상호간의 관계나 중요성 등을 연구하였다. 본 논문은 장대홍(1994)의 논문을 수정, 보완한

1) (608-737) 부산시 남구 대연 3동 599-1, 부산수산대학교 응용수학과 부교수.

것으로서, 이 biplot가 다반응값 혼합물 실험에서 뿐만이 아니라 다반응값 반응표면분석, 크게보면 다변량 회귀분석에도 쓰일 수 있음을 보이고자 하는데 이 논문의 목적이 있다.

2. 다변량 회귀분석을 위한 biplot

ξ_{ij}를 j번째 설명변수의 i번째 값 $(i=1,2,\cdots,n\ ;\ j=1,2,\cdots,k)$이라 하고, y_{ij}를 j번째 반응변수의 i번째 관측값 $(i=1,2,\cdots,n\ ;\ j=1,2,\cdots,m)$라 하자. 모든 반응변수들은 k개의 설명변수로 설명된다고 가정하고, 흥미영역에서 같은 차수(여기서는 1차)의 회귀모형을 갖는다고 하자. 같은 차수가 아니면 일반화 최소제곱추정방법을 사용하여야 한다.

ξ_{ij}와 y_{ij}를 표준화시키면,

$$x_{ij}=\frac{\xi_{ij}-\overline{\xi_j}}{\sqrt{\sum_{i=1}^{n}(\xi_{ij}-\overline{\xi_j})^2}}$$

이고,

$$Z_{ij}=\frac{y_{ij}-\overline{y_j}}{\sqrt{\sum_{i=1}^{n}(y_{ij}-\overline{y_j})^2}} \tag{1}$$

이 된다. 여기서, $\overline{\xi_j}=\frac{\xi_{ij}}{n}$ 이고, $\overline{y_j}=\sum_{i=1}^{n}\frac{y_{ij}}{n}$ 이다. 그러면, j번째 표준화된 반응변수에 대한 회귀식은 행렬을 이용하여

$$\underline{Z_j}=X\underline{\beta_j}+\underline{\varepsilon_j}\ ,\quad j=1,2,\cdots,m \tag{2}$$

로 표시할 수 있다. 여기서, X는 x_{ij}를 원소로 하는 n*k행렬이고, $\underline{\beta_j}$는 k*1벡터이다. 그래서, n*m 표준화 반응행렬(standardized response matrix) Z이 정의될 수 있는 데, $\underline{Z_j}$는 Z의 j번째 열이 된다. (2)식을 이용해 $\underline{\beta_j}$에 대한 추정값 $\underline{b_j}$를 구하면 다음과 같은 k*m 회귀계수행렬

$$B=[\ \underline{b_1}\ ,\ \underline{b_2}\ ,\ \cdots,\ \underline{b_m}]$$
$$=(X'X)^{-1}X'Z$$

이 구해진다. j번째 표준화된 반응변수의 추정값은

$$\hat{Z}_j(x)=b_1x_1+\cdots+b_kx_k$$

로 표시할 수 있는데, j번째 반응변수에 대한 i번째 설명변수 $(i=1,2,\cdots,k)$ 의 1차효과는 b_{ij}로

나타내어진다. 그리하여, B를 비정칙값 분해하여 biplot를 작성하면 반응변수들이나 설명변수들 상호간의 관계나 중요성등을 평가할 수 있게 된다. 여기서, 중요한 사실은 반응변수들이 1차 회귀모형을 갖지 못하고 2차 이상의 차수를 가지면, 교호작용항 때문에 j번째 반응변수에 대한 i번째 설명변수의 1차효과를 구별하여 내기가 어렵다는 것이다.

B의 rank를 r이라 할 때 $(r \leq \min(k, m))$ B에 대해 비정칙값 분해를 행하면

$$B = P\Lambda Q^T$$

이 된다. 여기서, P는 k*r행렬로서, P의 열들은 BB^T의 정규직교고유벡터(orthonormal eigenvector)들이고 Q^T는 r*m행렬로서, Q의 열들은 B^TB의 정규직교고유벡터들이다. Λ는 r*r 대각행렬로서, 대각요소가 비정칙값들인 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$이다.

이 B행렬을 대신해 rank-two행렬인 B^*로 만들면,

$$B^* = \lambda_1 \underline{p_1}\, \underline{q_1^T} + \lambda_2 \underline{p_2}\, \underline{q_2^T} = [\ \underline{p_1}, \underline{p_2}] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \underline{q_1^T} \\ \underline{q_2^T} \end{bmatrix}$$

가 된다. 여기서, $\underline{p_1}$과 $\underline{p_2}$는 각각 P 의 첫번째와 두번째 열이고, $\underline{q_1^T}$와 $\underline{q_2^T}$는 각각 Q^T의 첫번째와 두번째 행이다. B^*가 B를 대신할 수 있는지는, 기여율인

$$\rho_2 = \frac{\lambda_1^2 + \lambda_2^2}{\sum_{i=1}^{r} \lambda_i^2}$$

를 이용하여 점검할 수 있다. 대략 $\rho_2 > 0.9$이면 B^*로 B를 대신할 수 있다.

B^*를 요인분해(factorization)하는 방법에는 전형적으로 2가지가 있다. 즉, GH^T와 JK^T분해인데, GH^T분해는 B^*를

$$B^* = GH^T$$

로 분해하는 것이다. 여기서, k*2행렬 $G = [\ \underline{p_1}\ ,\ \underline{p_2}]$ 이고, m*2행렬 $H = [\ \lambda_1\underline{q_1}\ ,\ \lambda_2\underline{q_2}]$ 이다.

G의 처음부터 k번까지의 행들이 biplot의 행표시자 $\underline{r_1}, \underline{r_2}, \cdots, \underline{r_k}$가 되고 H의 처음부터 m번까지의 행들이 biplot의 열표시자 $\underline{c_1}, \underline{c_2}, \cdots, \underline{c_m}$가 된다. 이 biplot를 GH^T biplot라 부른다.

B^*의 요소를 b_{ij}^*라 하면

$$b_{ij}^* = \underline{g_i^T}\ \underline{h_j} = |\underline{g_i^T}|\,|\underline{h_j}|\cos(\theta_{ij})$$

로 표시할 수 있다. 여기서, $\underline{g_i^T}$는 G의 i번째 행이고, $\underline{h_j}$는 H^T의 j번째 열이다. 그러므로, i번째 설명변수와 j번째 반응변수 사이의 관계는, j번째 반응변수에 해당하는 벡터의 크기에, i번째 설명변수에 해당하는 벡터를 j번째 반응변수에 해당하는 벡터에 투영시켰을 때의 정사영의 크기를 곱

하면 나타나게 된다. B^*의 열들이 각각 요소들의 합이 0가 될 때는 $\underline{h_i}$와 $\underline{h_j}$벡터 사이의 cosine 값이 i번째 반응변수와 j번째 반응변수 사이의 상관계수의 근사값에 해당하게 된다. 그러므로, GH^T biplot를 통해서는 반응변수들간의 상관관계나 반응변수들과 설명변수들간의 관계를 규명할 수 있게 된다.

JK^T분해는

$$B^* = JK^T$$

로 분해하는 것이다. 여기서, k*2행렬 $J=[\ \lambda_1 \underline{p_1}\ ,\ \lambda_2 \underline{p_2}\ ]$ 이고, m*2행렬 $K=[\ \underline{q_1}\ ,\ \underline{q_2}\ ]$ 이다. J의 처음부터 k번까지의 행들이 biplot의 행표시자가 되고, K의 처음부터 m번까지의 행들이 biplot의 열표시자가 된다. JK^T대신 J에서의 행표시자들만 표시한 것이 j plot이다.

$$B^*B^{*T} = JJ^T$$

]이므로 J에서의 i번째 행표시자의 크기가, i번째 설명변수가 m개의 반응변수들에 끼치는 효과의 크기가 된다. 그러므로, j biplot를 통해서는 설명변수들의 중요도를 알아낼 수 있게 된다. B행렬을 대신하여 rank-two 행렬을 이용할 수 없다면 rank-three 행렬을 이용하여 GH^T plot 와 j plot를 작성하여야 하는데, 이 때, 이 biplot들은 3차원 그림이 된다.

Biplot를 작성하는 순서를 간략히 정리하면 다음과 같다.

1. 설명변수들과 반응변수들을 표준화시킨다.
2. 회귀식들을 구하여 B행렬을 만든다.
3. B행렬에 대하여 비정칙값 분해를 행한다.
4. B행렬을 대신하는 B^*행렬을 요인분해한다.
5. GH^T plot과 j plot를 그린다.

위의 순서들 중 3번의 비정칙값 분해가 가장 중요한 순서라 볼 수 있는데, 우리 주변에서 흔히 쓰이는 통계팩키지에는 비정칙값 분해를 행하는 함수들이 정의되어 있다. 예를들면, SAS/IML에서는 SVD라는 선형대수 함수가 있고, S-PLUS에서는 svd라는 함수가 있다.

3. 수치예

Khuri와 Conlon(1981)의 논문을 이용하여 아래와 같은 두 가지 예를 보이고자 한다.

예 1. 투석 유장단백질농축물(whey protein concentrates : WPC) 겔시스템의 조직 특성 연구에서 다음 표1과 같이 2개의 설명변수와 4개의 반응변수가 있다. 실험결과는 다음의 표2과 같고, 반응변수들을 (1)식을 이용해 표준화시킨 후 반응변수들에 대하여 1차 모형을 적용하여 회귀계수들을 구하였다. 표3은 B행렬, 1차모형을 적용하였을 때의 결정계수 R^2값들과 분산분석시의 p-값들을 나타낸 표이다. 두 번째 반응변수 y_2에 대한 분산분석시 p-값이 0.0753이 나왔으나 분석에는 문제가 되지않아 B행렬을 이용하여 biplot들을 작성하면 다음 그림1과 그림2와 같다.

표1. WPC 겔 시스템의 조직특성
연구에서의 변수들

설명변수	x_1	시스테인 (mM)
	x_2	염화칼슘 (mM)
반응변수	y_1	견고성 (kg)
	y_2	점착성
	y_3	탄력성 (mm)
	y_4	압축액 (g)

표2. 예1의 자료에 대한 실험결과

계획		반응값			
x_1	x_2	y_1	y_2	y_3	y_4
-1	-1	2.48	0.55	1.95	0.22
1	-1	0.91	0.52	1.37	0.67
-1	1	0.71	0.67	1.74	0.57
1	1	0.41	0.36	1.20	0.69
-1.414	0	2.28	0.59	1.75	0.33
1.414	0	0.35	0.31	1.13	0.67
0	-1.414	2.14	0.54	1.68	0.42
0	1.414	0.78	0.51	1.51	0.57
0	0	1.50	0.66	1.80	0.44
0	0	1.66	0.66	1.79	0.50
0	0	1.48	0.66	1.79	0.50
0	0	1.41	0.66	1.77	0.43
0	0	1.58	0.66	1.73	0.47

표 3. 예1의 자료에 대한 B행렬, R^2과 p-값

	y_1	y_2	y_3	y_4
x_1	-0.2393	-0.2233	-0.2867	0.2758
x_2	-0.2183	-0.0250	-0.0894	0.1528
R^2	0.84	0.40	0.72	0.80
p-값	0.0001	0.0753	0.0017	0.0004

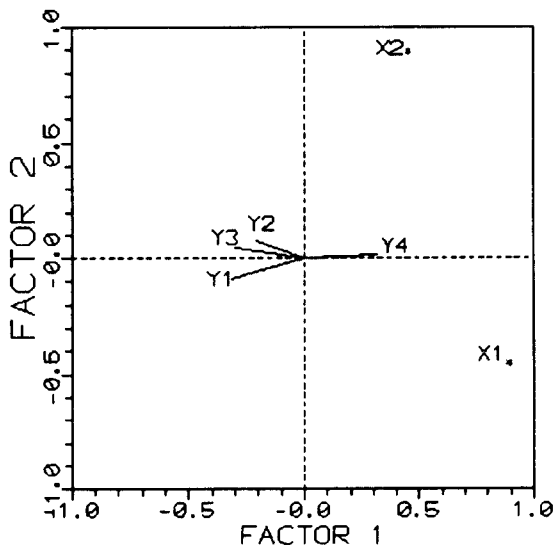

그림 1. 예 1의 자료에 대한 GH^T biplot

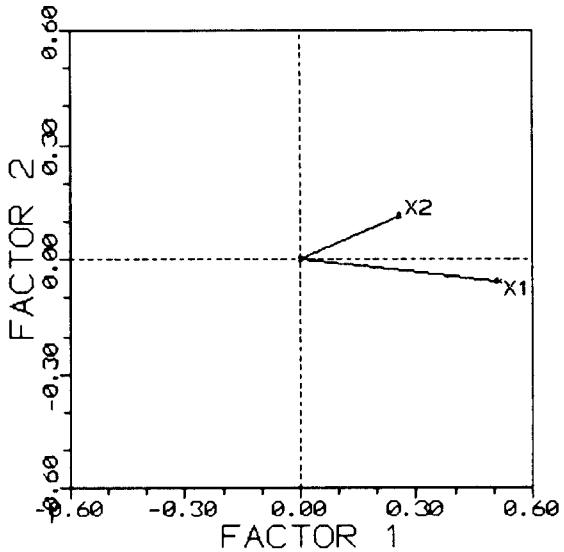

그림 2. 예 1의 자료에 대한 j biplot

그림1과 그림2를 통해 다음과 같은 사실을 알 수 있다.

1. 반응변수인 견고성, 점착성과 탄력성사이에는 양의 상관관계가 존재하고, 반응변수인 압축액은 앞의 3개의 반응변수들과 음의 상관관계가 있다.
2. 설명변수인 시스테인과 염화칼슘이 증가할수록 반응변수인 압축액이 증가하는 경향을 띈다. 영향을 끼치는 정도가 시스테인이 더 크다.
3. 설명변수인 시스테인과 염화칼슘이 증가할수록 견고성, 점착성, 탄력성이 감소하는 경향을 띈다. 영향을 끼치는 정도가 시스테인이 더 크다.
4. 설명변수의 효과가 시스테인 > 염화칼슘 순이다 (시스테인의 크기 : 0.5152, 염화칼슘의 크기 : 0.2822).

예 2. 유장단백질농축물 겔시스템의 거품성질 연구에서 다음의 표4와 같이 5개의 설명변수와 4개의 반응변수가 있다.

표4. WPC 겔 시스템의 거품성질 연구에서의 변수들.

설명변수	x_1	가열온도 (℃/30 min)
	x_2	PH 수준
	x_3	산화전위 (volt)
	x_4	수산화나트륨 (Molar)
	x_5	라우릴나트륨설페이트 (% of solids)
반응변수	y_1	최대범람 (%)
	y_2	최초낙하시간 (min)
	y_3	불변성단백질 (%)
	y_4	수용성단백질 (%)

실험결과는 표5와 같고, 반응변수들을 (1)식을 이용하여 표준화 시킨 후 반응변수들에 대하여 1차 모형을 적용하여 회귀계수들을 구하였다. 표6은 B행렬, 1차모형을 적용하였을 때의 결정계수R^2 값들과 분산분석시의 p-값들을 나타낸 표이다. 분산분석시 모두 유의하였고, B행렬을 이용하여 biplot를 작성하면 다음 그림3과 그림4와 같다.

표 5. 예2의 자료에 대한 실험결과

계획					반응값			
x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4
-1	-1	-1	-1	1	1082	4.5	80.6	81.4
1	-1	-1	-1	-1	824	7.5	67.9	69.6
-1	1	-1	-1	-1	953	8.3	83.1	105.0
1	1	-1	-1	1	759	17.0	38.1	81.2
-1	-1	1	-1	-1	1163	6.7	79.7	80.8
1	-1	1	-1	1	839	9.5	74.7	76.3
-1	1	1	-1	1	1343	12.0	71.2	103.0
1	1	1	-1	-1	736	36.0	36.8	76.9
-1	-1	-1	1	-1	1027	4.0	81.7	87.2
1	-1	-1	1	1	836	5.0	66.8	74.0
-1	1	-1	1	1	1272	12.5	73.0	98.5
1	1	-1	1	-1	825	20.0	40.5	94.1
-1	-1	1	1	1	1363	15.0	74.9	95.9
1	-1	1	1	-1	855	7.5	74.2	76.8
-1	1	1	1	-1	1284	18.5	63.5	100.0
1	1	1	1	1	851	12.0	42.8	104.0
-2	0	0	0	0	1283	12.0	80.9	100.0
2	0	0	0	0	651	8.5	42.4	50.5
0	-2	0	0	0	1217	4.5	73.4	71.2
0	2	0	0	0	982	10.5	45.0	101.0
0	0	-2	0	0	884	9.0	66.0	85.8
0	0	2	0	0	1147	9.0	71.7	103.0
0	0	0	-2	0	1081	9.0	77.5	104.0
0	0	0	2	0	1036	10.0	76.3	89.4
0	0	0	0	-2	1213	16.0	67.4	105.0
0	0	0	0	2	1103	8.5	86.5	113.0
0	0	0	0	0	1171	11.0	77.4	102.0
0	0	0	0	0	1179	9.0	74.6	104.0
0	0	0	0	0	1183	9.0	79.8	107.0
0	0	0	0	0	1120	10.0	78.3	104.0
0	0	0	0	0	1180	9.5	74.8	101.0
0	0	0	0	0	1195	11.0	80.9	103.0

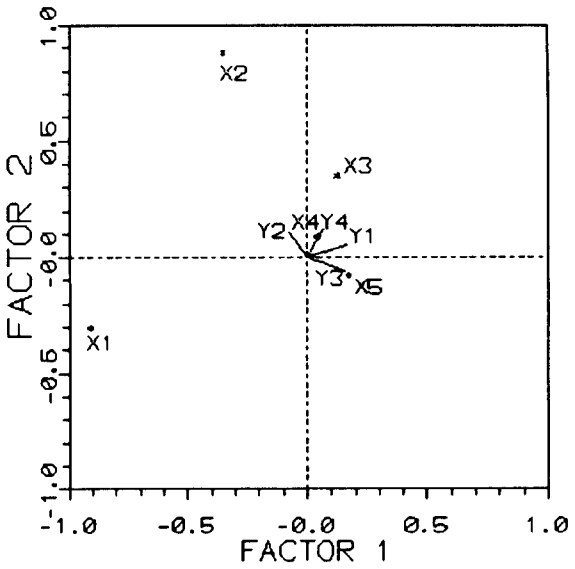

그림 3. 예2의 자료에 대한 GH^T biplot

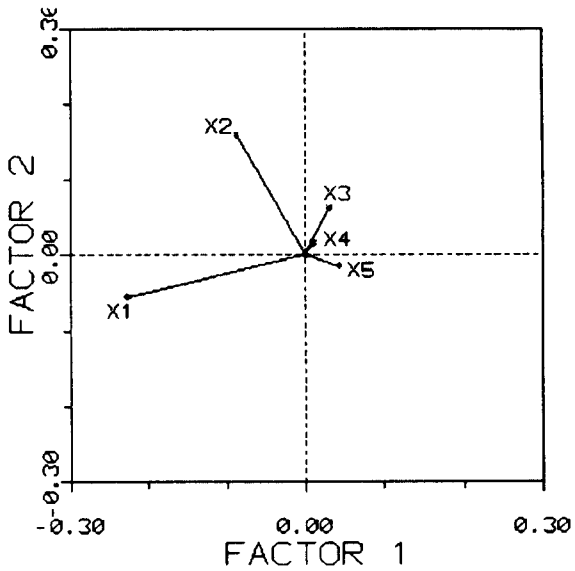

그림 4. 예 2의 자료에 대한 j biplot

표 6. 예2의 자료에 대한 B행렬, R^2과 p-값

	y_1	y_2	y_3	y_4
x_1	-0.1628	0.0326	-0.1281	-0.1022
x_2	-0.0168	0.1111	-0.1104	0.0932
x_3	0.0532	0.0482	0.0025	0.0295
x_4	0.0202	-0.0063	-0.0125	0.0140
x_5	0.0176	-0.0452	0.0206	0.0206
R^2	0.73	0.43	0.65	0.50
p-값	0.0001	0.0092	0.0001	0.0022

그림3과 그림4로 부터 다음과 같은 사실을 알 수 있다.

1. 설명변수인 가열온도가 높을수록 반응변수인 최초낙하시간은 늘어나고, 나머지 반응변수들은 줄어드는 경향을 띈다.
2. 설명변수들인 수산화나트륨과 라우릴나트륨설페이트는 반응변수들에 크게 영향을 끼치지 못한다.
3. 설명변수인 PH 수준이 높을수록 반응변수들인 최초낙하시간과 수용성단백질이 늘어나고, 불변성단백질은 줄어드는 경향이 있다.
4. 설명변수인 산화전위가 증가하면 최대범람, 최초낙하시간, 수용성단백질이 증가하는 경향이 있다.
5. 반응변수인 최대범람은 각각 불변성단백질, 수용성단백질과 양의 상관관계가 있으나, 불변성단백질과 수용성단백질과는 양의 상관관계가 크지 않다.
6. 최대범람과 최초낙하시간사이는 음의 상관관계가 크지 않다.
7. 최초낙하시간과 불변성단백질사이는 음의 상관관계가 있다.
8. 설명변수의 효과는 가열온도 > PH 수준 > 산화전위 > 라우릴나트륨설페이트 > 수산화나트륨순이다. 특히, 수산화나트륨의 효과는 미미하다(가열온도의 크기 : 0.2333, PH 수준의 크기 : 0.1825, 산화전위의 크기 : 0.0705, 라우릴나트륨설페이트의 크기 : 0.0456, 수산화나트륨의 크기 : 0.0193).

4. 결론

다반응값 반응표면분석에서 반응변수들이나 설명변수들 상호간의 관계나 중요성 등을 평가하는 것은 자료의 탐색적 단계에서 매우 중요하다. Biplot는 이러한 다반응값 반응표면분석에서 반응변수들 간의 상관관계와 설명변수들의 효과를 하나의 그림으로 나타낼 수 있는 유용한 도구이다. 설명변수들과 반응변수들을 표준화시킨 후 구한 회귀계수 추정값들을 이용하여 회귀계수행렬을 만든 후, 이 회귀계수행렬을 비정칙값 분해를 행하여 GH^T biplot와 j biplot를 작성할 수 있다. GH^T biplot를 통해서는 반응변수들 간의 상관관계나 반응변수들과 설명변수들 간의 관계를 규명할 수 있고, j biplot를 통해서는 설명변수들의 중요도를 알아낼 수 있다. 우리는 수치 예를 통하

여 다반응값 반응표면분석에 biplot가 유용하게 쓰일 수 있음을 보였다. 회귀모형이 1차 모형으로 적합치 않은 경우의 biplot 작성에 대한 연구가 차후의 작업이 될 것이다.

참고 문헌

[1] 장대홍(1994). Multiresponse Problem of Response Surface Methodology - biplot의 활용에 대한 연구, 회귀분석을 통한 품질향상에 관한 workshop(한국통계학회 공업통계연구회), 143-150.

[2] Corsten, L. C. A. and Gabriel, K. R.(1976). Graphical Exploration in Comparing Variance Matrices, *Biometrics*, 32, 851-863.

[3] Daigle, G. and Rivest, L. P.(1992). A Robust Biplot,*The Canadian Journal of Statistics*,Vol. 20, No. 3, 241-255.

[4] Gabriel, K. R.(1971). The Biplot-Graphic Display of Matrices with Application to Principle Component Analysis, *Biometrika*, 58, 453-467.

[5] Gabriel, K. R.(1978). Least Squares Approximation of Matrices by Additive and Multiplicative Models, *Journal of the Royal Statistical Society*, B, 40, 186-196.

[6] Gabriel, K. R.(1981). Biplot, *In Encyclopedia of Statistical Sciences, Vol. I, (S. Kotz, N. L. Johnson and C. Read, eds.).* New York, John Wiley, 262-265.

[7] Gabriel, K. R. and Odoroff, C. L.(1985). Illustration of Model Diagnosis by Means of Three-Dimensional Biplots, *Statistical Image Processing and Graphics, (E. J. Wegman and D. J. Depriest, eds.).* Marcel Dekker, 257-274.

[8] Gabriel, K. R. and Zamir, S.(1979). Lower Rank Approximation of Matrices by Least Sqraures with Any Choice of Weights, *Technometrics*, 21, 489-498.

[9] Gower, J. C.(1990). Three-dimensional Biplots, *Biometrika*, 77, 4, 773-785.

[10] Gower, J. C. and Harding, S. A.(1988). Nonlinear Biplots, *Biometrika*, 75, 3, 445-455.

[11] Michael Greenacre, J.(1993). Biplots in Correspondence Analysis, *Journal of Applied Statistics*, Vol. 20, No. 2, 251-269.

[12] Khuri, A. Z. and Conlon, M.(1981). Simultaneous Optimization of Multiple Responses Represented by Poynomial Regression Functions, *Technometrics*, Vol. 23, No. 4, 363-375.

[13] Smith, W. F. and Cornell, J. A.(1993). Biplot Displays for Looking at Multiple Response Data in Mixture Experiments, *Technometrics*, Vol. 35, No. 4, 337-350.