The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.083
30(1), 83–93

Visualizing multidimensional data in multiple groups

Myung-Hoe Huh[a,1]

[a]Department of Statistics, Korea University

(Received October 4, 2016; Revised November 28, 2016; Accepted November 29, 2016)

Abstract

A typical approach to visualizing $k\,(\geq 2)$-group multidimensional data is to use Fisher's canonical discriminant analysis (CDA). CDA finds the best low-dimensional subspace that accommodates k group centroids in the Mahalanobis space. This paper proposes an alternative visualization procedure functioning in the Euclidean space, which finds the primary dimension with maximum discrimination of k group centroids and the secondary dimension with maximum dispersion of all observational units. This hybrid procedure is especially useful when the number of groups k is two.

Keywords: canonical discriminant analysis, principal component analysis, biplot, Mahalanobis distance, scaled Euclidean distance

1. 연구 배경과 목적

이 연구가 상정하는 데이터는 $k\,(\geq 2)$ 그룹의 p-차원 연속형 관측들로서, 그룹 j의 관측 i를 $\boldsymbol{x}_{ij}$로 표기하고 $(j=1,\ldots,k;\ i=1,\ldots,n_j)$, $k<p$임을 가정한다 $(N=\sum_{j=1}^{k} n_j)$.

Fisher (1936)의 정준판별분석(canonical discriminant analysis; CDA)은 p개 변수 $X_1,\ldots,X_p$의 선형결합 $L = c_1X_1 + \cdots + c_pX_p$에서 그룹 중심점들 $\boldsymbol{c}^t\bar{\boldsymbol{x}}_1,\ldots,\boldsymbol{c}^t\bar{\boldsymbol{x}}_k$가 최대산포를 갖도록 $\boldsymbol{c} = (c_1,\ldots,c_p)^t$를 찾는다. 이 때, L의 그룹 내 분산들의 pooling을 1로 제약한다. 이에 따라 CDA는

$$\max_{\mathbf{c}} \frac{\sum_{j=1}^{k} n_j \left(\boldsymbol{c}^t\bar{\boldsymbol{x}}_{.j} - \boldsymbol{c}^t\bar{\boldsymbol{x}}_{..}\right)^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j} \left(\boldsymbol{c}^t\boldsymbol{x}_{ij} - \boldsymbol{c}^t\bar{\boldsymbol{x}}_{.j}\right)^2/(N-k)} \tag{1.1}$$

로 정식화된다. 이것은 수리적으로 다음과 같이 풀린다. B와 W를 각각 $p\times p$ 행렬

$$B = \sum_{j=1}^{k} n_j \left(\bar{\boldsymbol{x}}_{.j} - \bar{\boldsymbol{x}}_{..}\right)\left(\bar{\boldsymbol{x}}_{.j} - \bar{\boldsymbol{x}}_{..}\right)^t, \qquad W = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{.j})(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{.j})^t$$

로 정의하고 k개 그룹의 총합 공분산을 $S = W/(N-k)$라고 할 때, S로 정규화된 행렬 $B\,(= S^{-1/2}BS^{-1/2})$의 아이겐 분해(eigen-decomposition)에서 CDA의 해가 도출된다. 즉,

This research was supported by a Korea University Grant (K1609381).
[1]Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea. E-mail: stat420@korea.ac.kr

$$S^{-\frac{1}{2}}BS^{-\frac{1}{2}} = VD_\lambda V^t,$$

여기서 V는 직교하는 열로 구성된 고유벡터 행렬이고, D_λ는 $p \times p$ 대각행렬로서 대각선은 고유값으로 구성되는데 그 중 $k-1$개 값은 양이고 나머지는 0이다. 최적 $\boldsymbol{c}$의 해는 가장 큰 고유값 λ_1에 대응하는 V의 첫째 열 $\boldsymbol{v}_1$으로부터 얻어진다. 제1축 해의 구체적 수식 표현은 $\boldsymbol{c}_1 = S^{-1/2}\boldsymbol{v}_1$이고 이때 (1.1)의 값은 λ_1이다. $\boldsymbol{c}$의 최적 해를 $\boldsymbol{c}_1$으로 표기하자.

제2차원의 탐구는

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{\left(\boldsymbol{c}_1^t\boldsymbol{x}_{ij} - \boldsymbol{c}_1^t\bar{\boldsymbol{x}}_{\cdot j}\right)\left(\boldsymbol{c}^t\boldsymbol{x}_{ij} - \boldsymbol{c}^t\bar{\boldsymbol{x}}_{\cdot j}\right)}{(N-k)} = 0 \tag{1.2}$$

의 제약 하에서 식 (1.1)로 정식화된다. 여기서 제약 식 (1.2)는 $\boldsymbol{c}_1^t S\boldsymbol{c} = 0$이므로 마할라노비스 공간에서 $\boldsymbol{c}_1$과 $\boldsymbol{c}$가 직교함을 의미한다. 최적 $\boldsymbol{c}$의 해는 두 번째 큰 고유값 λ_2에 대응하는 V의 둘째 열 $\boldsymbol{v}_2$로부터 얻어진다. 즉, $\boldsymbol{c} = S^{-1/2}\boldsymbol{v}_2$이고 이것을 $\boldsymbol{c}_2$로 표기하자. 그때 (1.1)의 값은 λ_2이다.

이상의 방식을 거듭 적용하여 p-차원의 마할라노비스 공간에서 p개의 직교하는 정준 축을 찾을 수 있다. 이 중 $k'(=k-1)$개의 정준 축에서는 k개 그룹 중심점에 차이가 있지만 나머지 $p-k'$개의 정준 축에서 k개 그룹 중심점이 0에 위치하므로 그룹 변별의 의미가 없다.

$k=3$의 경우, 의미 있는 정준 축은 2개이므로 2차원 시각화에 잘 맞는다. 그러나 3개 그룹 중심점이 제1축에 가깝게 놓이는 경우 제2축은 아무 의미가 없는 나머지 축들과 별로 다르지 않을 수 있다. $k=2$의 경우, 의미 있는 정준 축은 1개이다. 다차원 데이터의 저차원 시각화에서 가장 편리한 차원 수가 2인 점을 감안하면 2차원 시각화 공간의 1개 정준 축은 그룹 변별에 활용하지만 남은 1개 정준 축은 활용하지 못하는 문제가 있다.

Iris 데이터를 사례로 보자. 이 데이터에서 그룹의 수 k는 3이고 측정변량의 차원 수 p는 4이다. Figure 1.1이 마할라노비스 공간의 2차원 시각화인데 화살표의 끝점은 변수 별로 1 표준편차 특성점의 사영이다. 이 사례에서 3개의 그룹(setosa, versicolor, virginica)의 중심점은 사실상 제1축에 놓인다. 식 (1.1)를 $(N-k)$로 나누어 신호대잡음 비(signal-to-noise ratio; SN ratio)

$$\frac{\sum_{j=1}^{k} n_j\left(\boldsymbol{c}^t\bar{\boldsymbol{x}}_{\cdot j} - \boldsymbol{c}^t\bar{\boldsymbol{x}}_{\cdot\cdot}\right)^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\boldsymbol{c}^t\boldsymbol{x}_{ij} - \boldsymbol{c}^t\bar{\boldsymbol{x}}_{\cdot j}\right)^2}$$

를 계산해보면, 제1축에서 32.2이고 제2축에서는 0.29로 1보다 훨씬 작다. 따라서 Figure 1.1에서 사영점들의 제2축 좌표는 거의 의미가 없다고 볼 수 있다. 예로서 Figure 1.2를 보자. 수직축에 정준판별분석의 제2축 좌표 대신 그룹 변별에서 전혀 역할이 없는 제4축 좌표를 넣은 것인데, Figure 1.1과 비교하여 그룹 분류에 있어 별 차이가 없다.

Iris 데이터에서 한 그룹(setosa)은 다른 두 그룹(versicolor, virginica)과 확연히 분리되어 있으므로 겹침이 있는 두 그룹 간 판별이 관건이다. 그런데 두 그룹 간 비교에서 정준판별분석은 1차원 시각화만 제시할 뿐이다. 이런 문제에 대한 인식에서 연구를 시작하였다.

다차원 데이터의 저차원 시각화 방법의 효시는 Gabriel (1971)이다. 그가 제안한 행렬도(biplot)는 주성분분석에서 행(관측)뿐만 아니라 열(변수)을 한 프레임에서 다루는 기하적 토대로 볼 수 있다. 이후, 방법론이 여러 방면으로 확대 발전하였다 (Gower와 Hand, 1996). 국내에서도 Choi 등 (2005), Huh (2013), Huh 등 (2007), Park와 Huh (1996)이 연구하였고 Choi와 Shin (2013)과 Huh (2012)에 정리되었다.

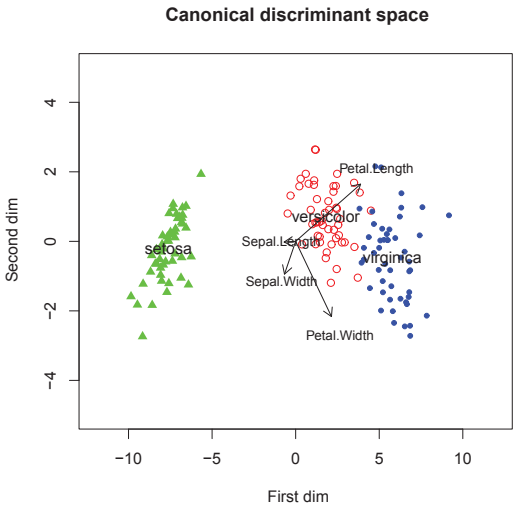

Figure 1.1. Visualization of the iris data in Mahalonobis space: the first and the second dimensions.

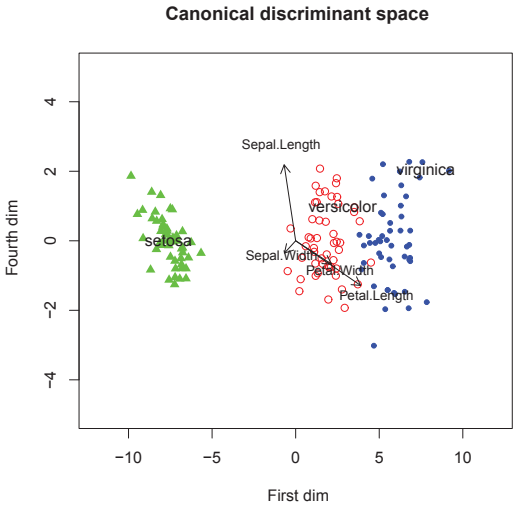

Figure 1.2. Visualization of the iris data in Mahalonobis space: the first and the fourth dimensions.

이 연구는 다그룹 다차원 데이터의 저차원 시각화에서 새로운 제2축을 제안한다. 마할라노비스 공간 대신 척도화 유클리드 공간(scaled Euclidean space)에서 작업할 것이다. 이러한 대안적 시각화가 기존의 CDA 시각화를 보완하여 다그룹 다차원 데이터에 대한 이해에 도움이 됨을 보이고자 한다.

2. 척도화 유클리드 공간에서의 정준판별 시각화

앞 절에서와 같이 p개 변수 $X_1, \ldots, X_p$의 선형결합 $L = c_1X_1 + \cdots + c_pX_p$에서 그룹 중심점들 $\boldsymbol{c}^t\bar{\boldsymbol{x}}_1, \ldots, \boldsymbol{c}^t\bar{\boldsymbol{x}}_k$가 최대산포를 갖도록 $\boldsymbol{c} = (c_1, \ldots, c_p)^t$를 찾는데 그 과정에서 계수벡터 $\boldsymbol{c}$에 대한 정규화

조건 $\boldsymbol{c}^t S\,\boldsymbol{c} = 1$을 $\boldsymbol{c}^t D_S \boldsymbol{c} = 1$로 대치하기로 한다. 여기서 D_S는 그룹 내 공분산행렬 S에서 대각선 벡터만 남긴 $p \times p$ 대각행렬이다. 그러면 식 (1.1)의 정식화는

$$\max_{\boldsymbol{c}} \boldsymbol{c}^t B \boldsymbol{c} \quad \text{subject to } \boldsymbol{c}^t D_S \boldsymbol{c} = 1 \tag{2.1}$$

로 바뀐다. 이것의 해는 D_S로 정규화된 행렬 $B(= D_S^{-1/2} B D_S^{-1/2})$의 아이겐 분해에서 나온다. 즉,

$$D_S^{-\frac{1}{2}} B D_S^{-\frac{1}{2}} = V D_\lambda V^t,$$

여기서 V는 직교하는 열로 구성된 고유벡터 행렬이고, D_λ는 $p \times p$ 대각행렬로서 대각선은 고유값으로 구성되는데 그 중 $k-1$개 값은 양이고 나머지는 0이다. 이후 과정은 1절에서와 같다.

척도화 유클리드 공간에서도 $k'(= k-1)$개의 정준 축에서는 k개 그룹 중심점에 차이가 있지만 나머지 $p - k'$개의 정준 축에서는 k개 그룹 중심점이 모두 0에 위치한다. 이제까지로 봐서는 척도화 유클리드 공간에서의 작업이 마할라노비스 공간에서의 작업과 별반 차이가 없다.

제1축 해는 정식화 (2.1)의 최대변별에서 찾되 제2축 해는 최대산포에서 찾아 관측점들을 사영하는 시각화 방법을 제안한다. 절차 및 계산은 다음과 같다.

정준판별 하이브리드 행렬도

1) 정식화 (2.1)은 데이터 행렬 $X\,(= (x_{uv}))$에 내적 표준화를 적용함으로써 간결하게 표현된다. 즉,

$$\tilde{x}_{uv} \quad \leftarrow \quad \frac{(x_{uv} - \bar{x}_{.v})}{s_v}$$

로 전(前)처리한다. 여기서 $s_1, \ldots, s_p$는 $X_1, \ldots, X_p$의 within-groups pooled standard deviation이다. 이런 전처리 후, (2.1)은

$$\max_{\boldsymbol{c}} \boldsymbol{c}^t \tilde{B}\,\boldsymbol{c} \quad \text{subject to } \boldsymbol{c}^t \boldsymbol{c} = 1 \tag{2.2}$$

로 표현된다. 여기서 $\tilde{B}$는 $\tilde{X}\,(= (\tilde{x}_{uv}))$로부터 산출되는 between-groups SS 행렬이다. 정식화 (2.2)로부터 제1축 기저벡터 $\boldsymbol{c}_1$을 산출한다.

2) $\tilde{\boldsymbol{x}}_{ij}$ $(i = 1, \ldots, n_i;\ j = 1, \ldots, k)$를 제1축 기저인 $\boldsymbol{c}_1$에 사영함으로써 발생하는 잔여부분 $\tilde{\boldsymbol{x}}_{ij} - \boldsymbol{c}_1^t \tilde{\boldsymbol{x}}_{ij} \boldsymbol{c}_1\,(= \tilde{\tilde{\boldsymbol{x}}}_{ij})$에 대하여 다음 최적화를 한다.

$$\max_{\boldsymbol{c}} \boldsymbol{c}^t \left(\sum_j \sum_i \tilde{\tilde{\boldsymbol{x}}}_{ij} \tilde{\tilde{\boldsymbol{x}}}_{ij}^t \right) \boldsymbol{c} \quad \text{subject to } \boldsymbol{c}^t \boldsymbol{c} = 1,$$

여기서 얻어지는 $\boldsymbol{c}\,(= \boldsymbol{c}_2)$는 $\boldsymbol{c}_1$과 직교한다. $\boldsymbol{c}_2$는 행렬 $\tilde{\tilde{X}}$의 제1 주성분 벡터이다.

이 알고리즘은 2차원 행렬도를 만들어내지만, 변별 최대화에 제1차원과 제2차원을 할당하고 산포 최대화에 제 3차원을 할당하는 경우 행렬도는 3차원이 된다. 3차원 행렬도에서는 단계 1에서 직교기저벡터 $\boldsymbol{c}_1$과 $\boldsymbol{c}_2$가 산출되고 단계 2에서 $\tilde{\boldsymbol{x}}_{ij}$를 $\boldsymbol{c}_1$과 $\boldsymbol{c}_2$에 사영하여 생성되는 잔여부분 $\tilde{\tilde{\boldsymbol{x}}}_{ij}$에 최적화를 적용하여 기저벡터 $\boldsymbol{c}_3$가 산출된다. 즉,

$$\tilde{\tilde{\boldsymbol{x}}}_{ij} = \tilde{\boldsymbol{x}}_{ij} - \boldsymbol{c}_1^t \tilde{\boldsymbol{x}}_{ij} \boldsymbol{c}_1 - \boldsymbol{c}_2^t \tilde{\boldsymbol{x}}_{ij} \boldsymbol{c}_2.$$

Figure 2.1이 iris 자료에 적용된 하이브리드 행렬도이다. 화살표의 끝점은 변수 별 1 표준편차 특성점

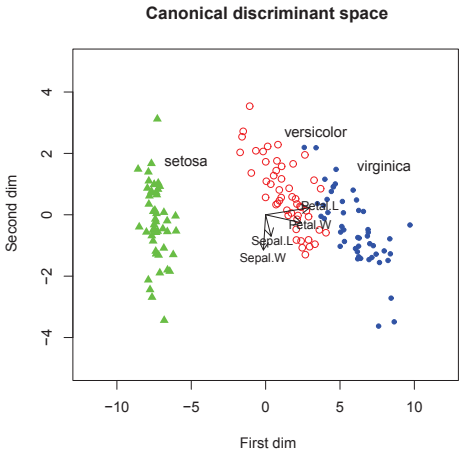

Figure 2.1. Visualization of the iris data in scaled Euclidean space: the first and the second dimensions .

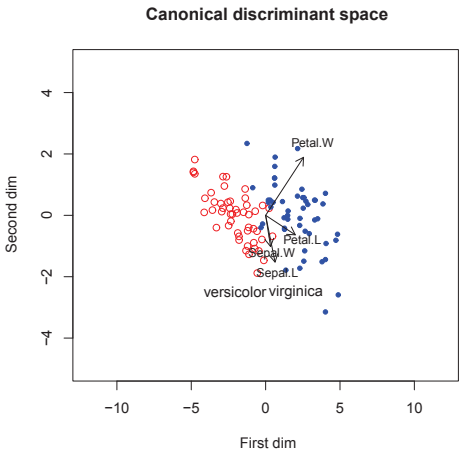

Figure 2.2. Visualization of the iris data in scaled Euclidean space: versicolor versus virginica.

의 사영이다. Iris 사례에서는 전통적인 정준판별분석과 특별히 다른 결과를 보이는 것은 아니다.

정준판별 하이브리드 행렬도 알고리즘은 그룹 수가 2인 경우에도 잘 작동한다. Figure 2.2는 iris 자료에서 versicolor 종과 virginica 종의 iris 자료 ($k = 2$, $N = 100$)에 대한 하이브리드 행렬도이다. 2개의 변수 petal length와 petal width가 종간 분류 정보를 적재한다(제1축). 반면 3개 변수 sepal length, sepal width, petal width의 결합과 petal length의 대비가 종내 산포를 가장 잘 보여준다(제2축).

3. 사례분석

3.1. Body 데이터

Gclus 팩키지의 body 데이터는 남자 247명과 여자 260명의 21개 신체부위와 나이, 체중, 신장 등으로 구성되어 있다. 성별로 나눈 21개 신체부위 관측 자료의 CDA 시각화를 해보자. 그룹 수 k가 2인 경우

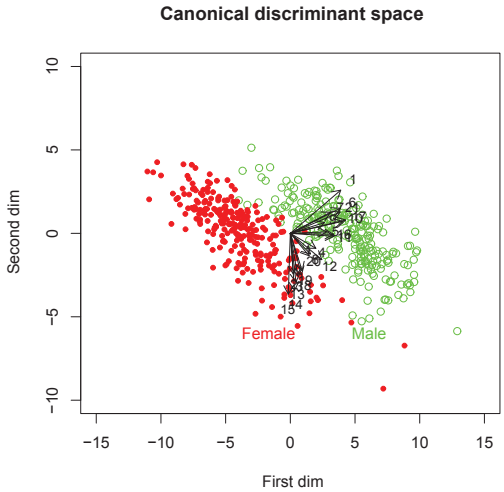

Figure 3.1. Visualization of the body data in scaled Euclidean space.

Table 3.1. Correlations between two component scores and three external variables

	Age	Weight	Height
Score 1	0.23	0.94	0.73
Score 2	−0.23	−0.66	−0.23

이므로 통상적인 정준판별분석의 2차원 시각화에서 제2축은 의미가 없다.

Figure 3.1는 척도화 유클리드 공간의 시각화인데, 이 그림에서 제1축은 그룹 간 변별의 최대화에서 나왔고 제2축은 산포 최대화로부터 나왔다. 이 그림은 성별 그룹이 타원체를 이루고 있음을 보여준다. Female 그룹에서 몇 개의 특이점이 발견된다.

Figure 3.1에서 제1축을 결정짓는 상위 5개 변수는 17: ForearmG, 10: ShoulderG, 1: Biacrom, 6: ElbowD, 21: WristG이며 모두 같은 방향이다. 제2축을 결정짓는 상위 4개 변수는 플러스 방향의 1: Biacrom (shoulder width)과 마이너스 방향의 15: ThighG, 14: HipG, 13: AbdG이다.

두 축의 성분점수 score 1, score 2와 외적변수인 age, weight, height 간 상관계수는 Table 3.1과 같다. Score 2와 weight 간에는 −0.66의 상관이 있다. 이것은 수직축의 플러스 쪽 관측점은 작은 체중, 마이너스 쪽 관측점은 큰 체중을 의미한다.

3.2. Olives 데이터

R 팩키지 classifly의 olives 데이터는 이탈리아의 3개 지역(North, South, Sardina)에서 수집된 올리브 유의 8개 지방산 측정값들로 구성되어 있다($n = 572$, $p = 8$, $k = 3$).

Figure 3.2는 마할라노비스 공간에서의 정준판별분석 시각화인데 제1축에 의하여 'region1'과 'region2 + region3'가 구분되고, 제2축에 의하여 'region2'와 'region3'가 구분되는 모습을 보인다. 제1축에서 가장 큰 영향이 있는 변수는 eicosenoic이고 제2축에서 가장 큰 영향이 있는 변수는 oleic이다.

Figure 3.3은 척도화 유클리드 공간에서의 하이브리드 행렬도인데 Figure 3.2와는 다른 데이터 특징이 포착된다. 예컨대, 그룹 별 자료점 클라우드가 모두 말발굽(horse shoe) 모습을 하고 있는데 이것은

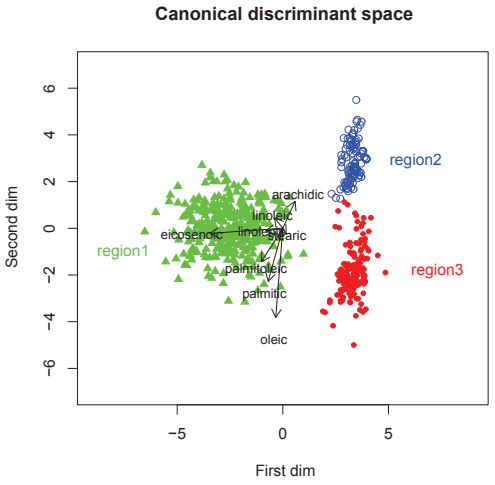

Figure 3.2. Visualization of the olives data in Mahalanobis space.

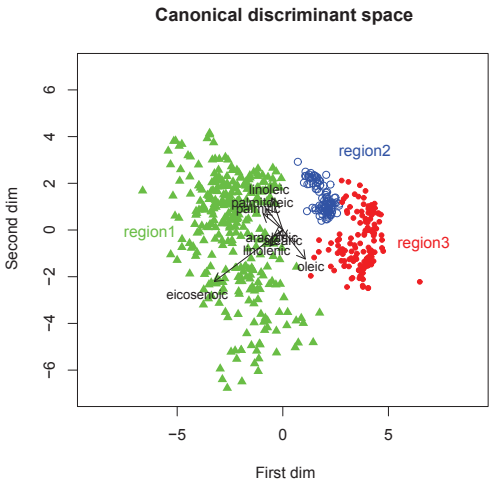

Figure 3.3. Visualization of the olives data in scaled Euclidean space.

Figure 3.1에서는 볼 수 없었던 점이다.

앞에서 관찰하였듯이 Figure 3.2에서 oleic 변수는 매우 중요하다. Table 3.2에서 상관계수행렬을 조회해보면 oleic ($= v4$)과 palmitic ($= v1$) 간 상관계수는 -0.84로 음의 상관이 크다. 그렇지만 그림에서 두 변수는 유사한 방향을 지향하는 것으로 나타나 있다. 이에 반하여 Figure 3.3에서 두 변수는 반대 방향을 지향한다. oleic ($= v4$)과 palmitoleic ($= v2$) 간 관계도 Figure 3.2와 Figure 3.3에서 판이하게 다르다. 어느 이야기를 따라야 할까? 다음 절에서 이런 이슈를 다룰 것이다.

4. 마할라노비스 공간과 척도화 유클리드 공간에서의 변수 벡터

이 절에서는 2-변량 관측 (X_1, X_2)에 대한 마할라노비스 공간과 척도화 유클리드 공간을 생각하기로 한

Table 3.2. Correlations among variables of the olives data

	$v1$	$v2$	$v3$	$v4$	$v5$	$v6$	$v7$	$v8$
$v1$: Palmitic	1.00	0.84	−0.17	−0.84	0.46	0.32	0.23	0.50
$v2$: Palmitoleic	0.84	1.00	−0.22	−0.85	0.62	0.09	0.09	0.42
$v3$: Stearic	−0.17	−0.22	1.00	0.11	−0.20	0.02	−0.04	0.14
$v4$: Oleic	−0.84	−0.85	0.11	1.00	−0.85	−0.22	−0.32	−0.42
$v5$: Linoleic	0.46	0.62	−0.20	−0.85	1.00	−0.06	0.21	0.09
$v6$: Linolenic	0.32	0.09	0.02	−0.22	−0.06	1.00	0.62	0.58
$v7$: Arachidic	0.23	0.09	−0.04	−0.32	0.21	0.62	1.00	0.33
$v8$: Eicosenoic	0.50	0.42	0.14	−0.42	0.09	0.58	0.33	1.00

다. 그리고 (X_1, X_2)에 대하여 다음의 공분산 구조를 상정한다.

$$\sum = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \rho > 0.$$

대칭행렬 Σ에 대한 스펙트럼 분해는

$$\sum = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \tag{4.1}$$

이다. 따라서 유클리드 공간에서 2개의 주성분 점수 축은

$$S_1 = \frac{(X_1 + X_2)}{\sqrt{2}}, \quad S_2 = \frac{(X_1 - X_2)}{\sqrt{2}}$$

에 의하여 결정된다. 이에 따라 변수 특성점은 (S_1, S_2) 공간에

$$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \quad \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$$

에 타점된다. 즉, 2개의 변수벡터는 직교한다. $\rho < 0$인 경우엔 2개의 주성분 점수 축은

$$S_1 = \frac{(X_1 - X_2)}{\sqrt{2}}, \quad S_2 = \frac{(X_1 + X_2)}{\sqrt{2}}$$

이고 변수 특성점은 (S_1, S_2) 공간에

$$\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right), \quad \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$$

에 타점된다. 이들 벡터는 직교한다.

한편, $\rho > 0$인 경우, 마할라노비스 공간에서 2개의 주성분 점수는 식 (4.1)로부터

$$S_1 = \frac{(X_1 + X_2)}{\sqrt{2(1+\rho)}}, \quad S_2 = \frac{(X_1 - X_2)}{\sqrt{2(1-\rho)}}$$

로 결정되므로 변수 특성점은 (S_1, S_2) 공간에

$$\left(\frac{1}{\sqrt{2(1+\rho)}}, \frac{1}{\sqrt{2(1-\rho)}}\right), \quad \left(\frac{1}{\sqrt{2(1+\rho)}}, -\frac{1}{\sqrt{2(1-\rho)}}\right)$$

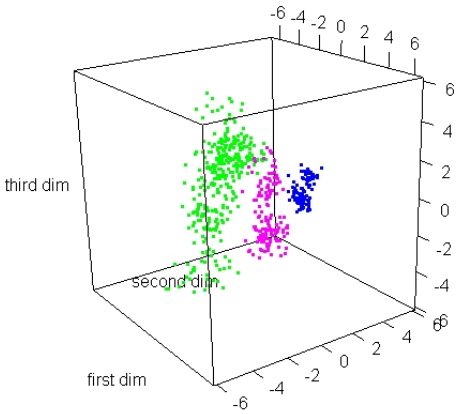

Figure 5.1. Visualization of the olives data in scaled Euclidean space: three dimensions.

에 타점된다. 따라서 2개의 변수벡터 간 내적은 $-\rho/(1-\rho^2)$로 음이 된다($\rho > 0$이므로). 즉, 2개 변수벡터 간 사이 각은 $\pi/2$보다 크게 나타난다. $\rho < 0$인 경우에는 변수 특성점은 (S_1, S_2) 공간에

$$\left(\frac{1}{\sqrt{2(1-\rho)}}, \frac{1}{\sqrt{2(1+\rho)}}\right), \quad \left(\frac{1}{\sqrt{2(1-\rho)}}, -\frac{1}{\sqrt{2(1+\rho)}}\right)$$

에 타점되고 2개의 변수벡터 간 내적은 $\rho/(1-\rho^2)$로 양이므로, 2개 변수벡터 간 사이 각은 $\pi/2$보다 작게 나타난다.

이상을 정리하면, 마할라노비스 공간에서는 2개 양상관 변수의 지향 화살은 둔각을 이루고 2개 음상관 변수의 지향 화살은 예각을 이룬다. 이렇듯 마할라노비스 거리가 적용되는 CDA 시각화는 통상적인 직관과 상충한다. 이런 이유로, 정준판별분석에서 마할라노비스 거리 대신 척도화 유클리드 거리를 대안으로 고려할 필요가 있다.

5. 맺음말

저차원 공간 시각화에서 차원 수는 통상 2이지만 R 환경에서는 3도 무리가 없는 선택이다. Rgl 팩키지의 plot3d 함수를 써서 3차원 산점도의 동적 그래픽스를 구현할 수 있기 때문이다. 분석 데이터에서 그룹 수 k가 3인 경우, 그룹 변별을 위해 제1축과 제2축을 쓰고 산포 최대화에 제3축을 쓸 것을 제안한다. Figure 5.1이 olives 자료에 적용된 척도화 유클리드 공간 시각화의 보기이다.

추기

이 논문에서 사용된 R 스크립트는 저자에 e-메일로 요청하여 받을 수 있다.

References

Choi, Y. S., Hyun, G. H., and Jung, S. M. (2005). MANCOVA biplot, *Korean Communications in Statistics*, **12**, 705–712.

Choi, Y. S. and Shin, S. M. (2013). *Understanding Biplots Analysis Using R*, Freedom Academy, Korea.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179–188.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453–467.

Gower, J. C. and Hand, D. J. (1996). *Biplots*, Chapman and Hall, London.

Huh, M. H. (2012). *Exploratory Multivariate Data Analysis*, Freedom Academy, Korea.

Huh, M. H. (2013). Biplots of multivariate data guided by linear and/or logistic regression, *Communications for Statistical Applications and Methods*, **20**, 129–136.

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

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

다그룹 다차원 데이터의 시각화

허명회[a,1]

[a]고려대학교 통계학과

(2016년 10월 4일 접수, 2016년 11월 28일 수정, 2016년 11월 29일 채택)

요 약

$k\,(\geq 2)$ 그룹의 p-차원 데이터의 시각화에서 가장 전형적인 방법은 Fisher의 정준판별분석(canonical discriminant analysis; CDA)이다. CDA는 마할라노비스 공간에서 k개 그룹 중심을 근사하게 통과하는 저차원 부공간에 관측점들을 사영한다. 본 논문은 척도화 유클리드 공간에서 다그룹 다차원 데이터를 시각화하는 방법을 제안하는데, 저차원 부공간의 제1축(또는 제1축과 제2축)은 그룹 중심들의 최대변별(maximum discrimination)에서 찾고 부공간의 제2축(또는 제3축)은 관측개체들의 최대산포(maximum dispersion)에서 찾는다. 이러한 혼종방법(hybrid method)은 2-그룹 다차원 자료의 시각화에서 특히 유용하다.

주요용어: 정준판별분석, 주성분분석, 행렬도(biplot), 마할라노비스 거리, 척도화 유클리드 거리

이 연구는 고려대학교 특별연구비에 의하여 수행되었음(K1609381).

[1](02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: stat420@korea.ac.kr