The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.4.555
30(4), 555–566

Representing variables in the latent space

Myung-Hoe Huh[a,1]

[a]Department of Statistics, Korea University

(Received May 8, 2017; Revised June 10, 2017; Accepted June 10, 2017)

Abstract

For multivariate datasets with large number of variables, classical dimensional reduction methods such as principal component analysis may not be effective for data visualization. The underlying reason is that the dimensionality of the space of variables is often larger than two or three, while the visualization to the human eye is most effective with two or three dimensions. This paper proposes a working procedure which first partitions the variables into several "latent" clusters, explores individual data subsets, and finally integrates findings. We use R pakacage "ClustOfVar" for partitioning variables around latent dimensions and the principal component biplot method to visualize within-cluster patterns. Additionally, we use the technique for embedding supplementary variables to figure out the relationships between within-cluster variables and outside variables.

Keywords: data visualization, clustering of variables, latent variables, principal component analysis, biplot, supplementary variables

1. 연구 배경과 목적

다변량 연속형 데이터에서 주성분분석(principal component analysis) 행렬도(biplot)는 관측개체들과 변수들을 저차원 공간에 시각적으로 동시에 표현해준다 (Gabriel, 1971). 이때 저차원 공간의 차원은 2–3 정도여야 하는데 차원이 그것보다 크게 되면 인간의 인지적 한계를 넘기 때문이다. 이는 분석 데이터의 p개 변수들에 대한 잠재변인(latent variable)이 2–3개를 넘지 않는 경우에만 주성분분석이 효과적임을 의미한다. 분석변수의 수 p가 10을 훨씬 초과하는 데이터들에서는 기대하기 어려운 조건이다.

이 연구의 목표는 잠재적 차원이 3 이상일 수 있는 다변량 데이터의 행(관측개체)과 열(변수)을 담은 유용한 행렬도를 개발하는 데 있다. 이를 위해 다음의 방법을 제안한다.

첫째, p개 변수들을 K개 그룹으로 잠재적 군집화한다 ($K \ll p$). 변수군집화에는 몇 가지가 있지만 군집 내 변수들을 잠재변인으로 엮어내는 "잠재변인 변수군집화" (이하 "잠재적 변수군집화"로 약칭) 방법을 쓸 것이다.

둘째, 개별 변수군집의 주성분분석 행렬도로 군집 내 변수들과 관측개체들을 시각화한다. 이때 행렬도의 차원 수는 편의상 2로 둔다. 따라서 행렬도의 제1축이 해당 변수군집을 대표하는 잠재변인이다. 행렬도의 제2축은 군집 내 변수들의 개별성을 탐구하는 데 쓴다. 개별 변수군집 행렬도에 외부 잠재변인

[1]Department of Statistics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
E-mail: stat420@korea.ac.kr.

들 또는 외적 변수를 “추가변수(supplementary variable)”로 끼워 넣어 군집 내 변수들과의 관계를 탐구한다.

논문의 2절에서 잠재적 변수군집화 방법을 설명하고, 3절에서 “추가변수 끼워넣기(embedding supplementary variables)” 기법의 기하를 정립할 것이다. 4절에서 몇 개의 데이터 사례를 다룬다.

2. 잠재적 변수군집화

데이터가 p개 변량의 n개 개체로 구성됨을 전제하고 p개의 길이-n 변수벡터를 $\mathbf{x}_1, \ldots, \mathbf{x}_p$로 표기하자. 특별한 언급이 없는 경우 모든 변량이 연속형이라고 가정한다.

잠재적 변수군집화의 목표는 p개 변량을 K개 잠재변인 중심으로 분할하는데 있다. 임의의 분할을 $P_K = (C_1, \ldots, C_K)$로 표기하자. R 팩키지 ClustOfVar과 ClustVarLV에 잠재적 변수군집화 방법이 구현되어 있다 (Chavent *et al.*, 2012, 2013; Vigneau *et al.*, 2015).

ClustOfVar에서 변수군집 $C_k\,(k = 1, \ldots, K)$를 대표하는 잠재변인 $\mathbf{u}_k$는

$$\mathbf{u}_k = \arg\max_{\mathbf{u}\in R^n} \sum_{c\in C_k} \mathrm{corr}^2(\mathbf{u}, \mathbf{x}_j) \tag{2.1}$$

이며, $\mathbf{u}_k$는 변수군집 C_k의 제1주성분 점수와 일치한다.

잠재변인 $\mathbf{u}_k$가 정해지면 군집 C_k의 “균질도(subgroup homogeneity)” $H(C_k)$를 산출한다:

$$H(C_k) = \sum_{\mathbf{x}_j\in C_k} \mathrm{corr}^2(\mathbf{u_k}, \mathbf{x}_j).$$

변수 $\mathbf{x}_j$가 범주형인 경우에는 앞의 두 수식에서 제곱상관이 상관비(correlaltion ratio)

$$\eta^2_{\mathbf{u},\mathbf{x}_j} = \frac{\sum_{l=1}^{L} n_l(\bar{\mathbf{u}}_l - \bar{\mathbf{u}})^2}{\sum_{i=1}^{n}(u_i - \bar{\mathbf{u}})^2}$$

로 대치된다. 여기서 l은 개별 범주를 지칭하고 L은 범주의 수, n_l는 범주 l내 개체들의 수이다. 그리고 $\bar{\mathbf{u}}_l$은 범주 l내 개체들의 u_i의 평균, $\bar{\mathbf{u}}$는 모든 개체들의 u_i의 평균이다.

전체변수의 분할 $P_K = (C_1, \ldots, C_K)$의 총 균질도(total homogeneity)는

$$H(P_K) = \sum_{k=1}^{K} H(C_k)$$

로 정의된다. 변수집합의 최적 분할은 계층적 군집화(hierarchical clustering), 또는 일종의 k-평균 군집화로 얻어낸다.

ClustVarLV에 구현된 방법은 앞의 (2.1) 또는

$$\mathbf{u}_k = \arg\max_{\mathbf{u}\in R^n} \sum_{\mathbf{x}_j\in C_k} \mathrm{corr}(\mathbf{u}, \mathbf{x}_j) \tag{2.2}$$

로부터 잠재변인 $\mathbf{u}_k$를 산출하기도 한다. 어떤 상황에서는, n차원 초구(hyper-sphere) 표면에서 지역적 인근성을 고려하는 것이 맞다는 취지이다 (Vigneau와 Quannari, 2003).

ClustOfVar와 ClustVarLV는 범주형 변수를 다루는 데 있어서 차이가 있다. 전자는 각 범주형 변수를 1개 차원으로 간주하는 데 비해 후자는 각 범주형 변수를 “범주의 수”만큼의 독자적 변수들로 간주한다. 따라서 ClustVarLV에서는 1개의 범주형 변수가 범주별로 다른 변수군집에 할당될 수 있다.

변수군집 수 K를 정하는 방법에서도 ClustOfVar와 ClustVarLV가 다르다. ClustOfVar은 adjusted Rand index로 군집화의 안정성(stability)을 평가하여 최대로 안정적인 K 값을 찾는다. 반면, ClustVarLV는 군집화 지표의 변화를 관찰하여 총 분산 중 설명되지 않은 부분이 급격히 증가하기 직전에 변수군집의 분할을 멈춘다.

이 연구에서는 ClustOfVar로 잠재적 변수군집화를 할 것이다. 변수군집 수를 정하는 데 있어 상대적으로 명확한 절차가 있다는 것 외에 특별한 이유가 있는 것은 아니다.

3. 주성분분석 행렬도에서의 추가변수 표현

다변량 데이터를 축소차원에 시각화하는 기본적인 방법은 주성분분석이다. n개의 p변량 데이터를 $n \times p$ 행렬 X로 표기하고 열 표준화가 적용되었다고 하자. X를 고유분해하여

$$X = UD_\mu V^t$$

를 산출한다. q-차원 주성분분석 행렬도는 n개 관측(observations)과 p개 변수(variables)를 다음 행렬의 행에서 취하여 각각 데이터 점과 원점에서 출발하는 화살의 종착점으로 나타낸 그래프이다 ($q \ll p$).

$$n \text{ observations} : U_{(q)}D_{\mu(q)}, \qquad p \text{ variables} : V_{(q)}.$$

여기서 $U_{(q)}$는 U의 첫 q개 열로 이루어진 $n \times q$ 부행렬이고 $D_{\mu(q)}$는 D_μ의 좌상 $q \times q$ 부행렬, 그리고 V_q는 V의 첫 q개 열로 이루어진 $p \times q$ 부행렬이다. 이 논문에서는 효과적인 시각적 전달을 위하여, 특별한 언급이 없는 경우, 개별변수 및 잠재변인의 화살 길이를 3배로 늘려 잡았다.

이 그래프에 p^*개의 추가변수를 넣어보자. 추가변수는 그래프 작성에 포함되지 않은 변수를 일컫는데, 이런 변수들의 끼워넣기(embedding)는 기존 그래프에 영향을 주지 않는다는 제한 하에서 추가변수와 기존 분석변수들간의 관계를 시각적으로 드러내는 데 목표가 있다.

추가 데이터를 X^* ($n \times p^*$ 행렬)로 표기할 때

1) 그래프에서 X^t의 p개 행이 $V_{(q)}$의 p개 행에 매핑되므로
2) 그 선형변환은 $(XX^t)^- XV_{(q)}$로 결정된다. 따라서 n-차원 벡터 $\mathbf{x}^*$를 화살(arrow)로 표현할 때 화살의 출발점은 원점이고 종착점은

$$\mathbf{x}^{*t}(XX^t)^- XV_{(q)}$$

이다. 여기서 A^-는 A의 Moore-Penrose 일반화 역행렬의 표기이다.

이에 따라 X^*의 p^*개 열은 q-차원 행렬도에서 $X^{*t}(XX^t)^- XV_{(q)}$에 타점된다. 그런데

$$\begin{aligned} XX^t &= UD_\mu V^t V D_\mu U^t = UD_\mu^2 U^t, \\ (XX^t)^- &= UD_\mu^{-2} U^t \end{aligned}$$

이므로

$$X^{*t}(XX^t)^- XV_{(q)} = X^{*t}UD_\mu^{-2}U^t UD_\mu V^t V_{(q)} = X^{*t}U_{(q)}D_{\mu(q)}^{-1} \tag{3.1}$$

이다. 그러므로 X^{*t}의 j^* 열 벡터 $\mathbf{x}_j^* = (x_{1j^*}, \ldots, x_{nj^*})^t$는 q-차원 행렬도의

$$\left(\sum_{i=1}^n \frac{x_{1j^*}u_{i1}}{\mu_1}, \ldots, \sum_{i=1}^n \frac{x_{ij^*}u_{ik}}{\mu_k} \right) \tag{3.2}$$

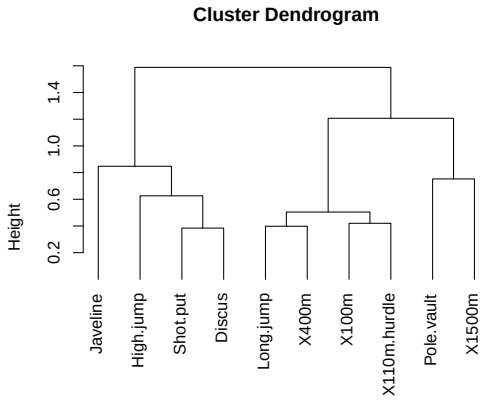

Figure 4.1. Dendrogram for variable clustering of the Decathlon data.

에 매핑 된다. 식 (3.1)과 동치인 식 (3.2)는 전환공식(transition formula) 또는 barycentric formulae로 불리면서 이 분야의 연구가 활발한 유럽에서 행 수량화와 열 수량화 결과 간 관련성을 일반화하는 방식으로 활용되고 있다 (Benzecri, 1992).

식 (3.1) 또는 식 (3.2)를 활용하여 다음과 같은 시각화 작업을 할 수 있다.

1) 개별 변수군집의 주성분분석 행렬도를 만들어 외부 군집의 잠재변인들을 끼워넣기.

2) 그 그래프에 외적 변수를 끼워넣기.

다음 절에서 실제 사례에 적용해 보기로 한다.

4. 사례

4.1. Decathlon

이 사례의 자료는 R 팩키지 FactoMineR에 포함되어 있다. 분석목표는 'Decastar' 또는 'OlympicG' 게임에서 선수 41명의 경기 종목 100m, Long.jump, Shot.put, High.jump, 400m, 110m.hurdle, Discus, Pole.vault, Javeline, 1500m 기록에서 10개 종목이 몇 개의 신체적 능력에 의해 결정되는지를 탐구하는 데 있다. 값이 클수록 좋은 경기결과를 나타내도록 시간으로 측정되는 100m, 400m, 110m.hurdle, 1500m에 대하여는 자료 값의 부호를 음으로 바꾸었다.

R 팩키지 ClustOfVar로 얻은 군집 덴드로그램은 Figure 4.1과 같다. 안정성 평가를 통해 3–6개의 변수군집이 적정한 것으로 보였다. 여기서는 10개의 변수를 3개 군집으로 나누기로 한다. 제 1군집은 100m, 110m.hurdle, 400m, Long.jump이며 제 2군집은 Shot.put, High.jump, Discus, Javeline, 제 3군집은 Pole.jump와 1500m가 되었다. 제 1잠재변인은 "뛰기 능력", 제 2잠재변인은 주로 "던지기 능력"이라고 할 수 있고, 제 3잠재변인은 "그 외의 능력"으로 해석할 수 있다.

Figure 4.2에서 변수군집 별로 잠재적 구조를 살펴보자. 개별 그래프는 변수군집별 2차원 행렬도이므로 이중 첫째 부차원(subdim 1)은 잠재차원(변인)과 일치한다. 그리고 그래프에서 "Dim 1", "Dim 2", "Dim 3" 화살은 제 1잠재변인, 제 1잠재변인, 제 2잠재변인, 제 3잠재변인을 지칭한다.

제 1변수군집 행렬도에서 제 1잠재변인과 제2잠재변인 간 약하게 음적인 관계가 있음을 볼 수 있다. 즉 "뛰기 능력"은 "던지기 능력"과 상반된다. 제 2변수군집 행렬도에서는 Discus(원반)와 High.jump(높

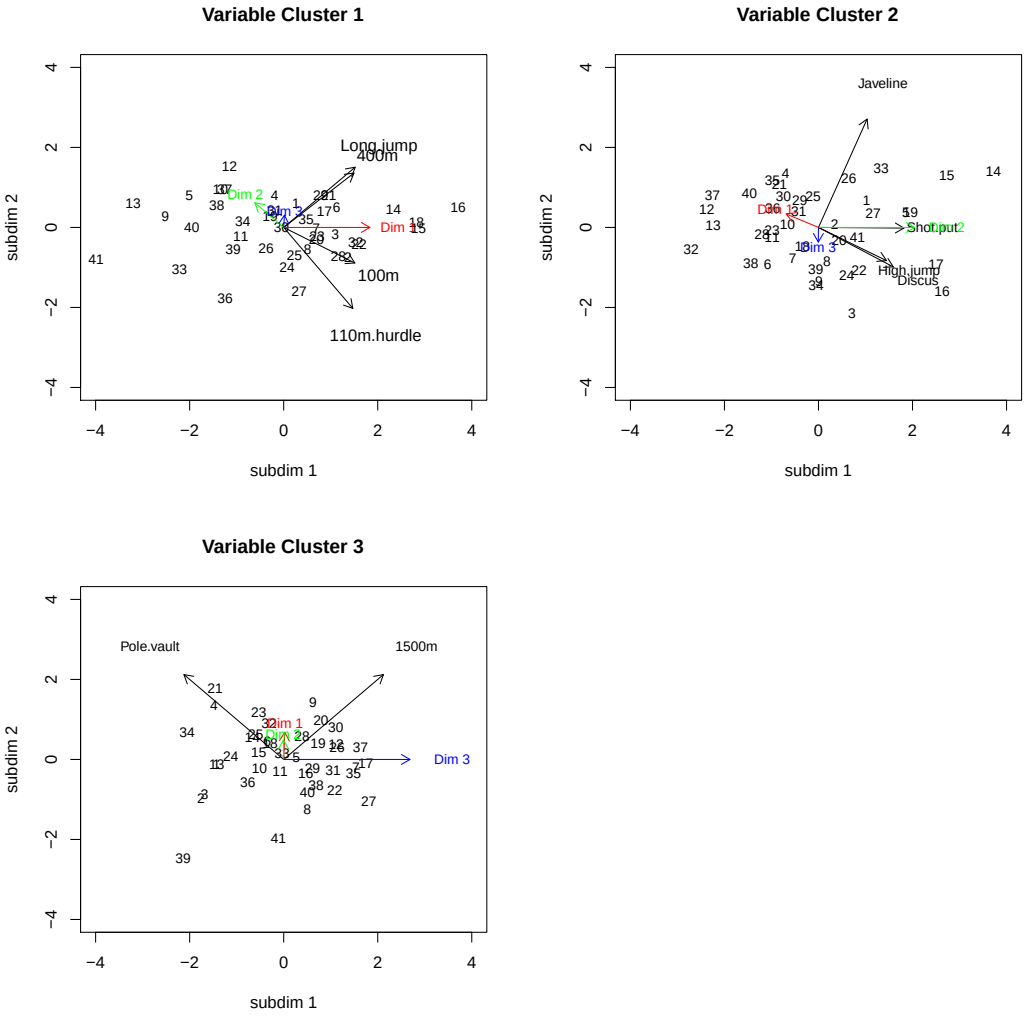

Figure 4.2. Biplots of the Decathlon data by variable clusters.

이뛰기)가 제 1잠재변인과의 역 상관성을 보여준다. 제 1잠재변인과 제 2잠재변인은 각각 제 3잠재변인과 독립적인 관계에 있는 것으로 보인다. 제 3변수군집에서 1500m와 Pole.vault가 묶여있지만 두 변수의 방향이 다르게 드러났다. 이것은 두 종목을 모두 잘하기는 어려움을 의미한다.

4.2. Body parts

gclus 팩키지의 body 데이터는 남자 247명, 여자 260명에 대한 21개 신체부위와 Age, Weight, Height 등의 정보를 담고 있다. 여기서는 남자 247명의 21개 신체부위를 몇 개의 잠재변인 변수군집으로 나누고자 한다.

ClustOfVar 팩키지를 써서 변수군집화를 해보면 5–7개의 변수군집이 가장 안정성이 있는 것으로 나타나는데, 여기서는 5개를 취하기로 한다. Figure 4.3에서 덴드로그램을 보라. 이 사례처럼 변수군집 수

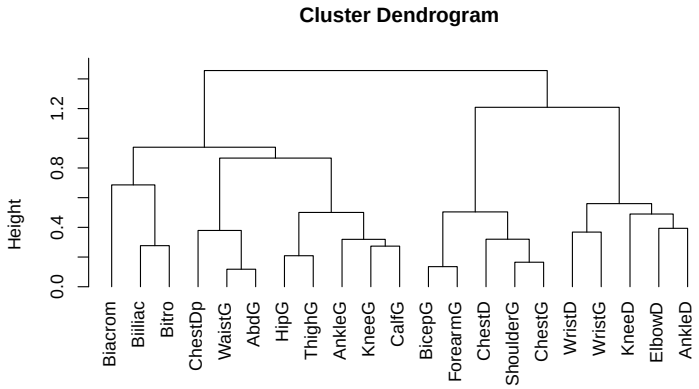

Figure 4.3. Dendrogram for variable clustering of the body parts data.

Table 4.1. Correlation matrix of the body parts data

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	1.00	−0.56	−0.49	−0.53	−0.64
Cluster 2	−0.56	1.00	0.68	0.46	0.70
Cluster 3	−0.49	0.68	1.00	0.65	0.71
Cluster 4	−0.53	0.46	0.65	1.00	0.65
Cluster 5	−0.64	0.70	0.71	0.65	1.00

가 3을 넘어서는 경우 기존의 주성분분석은 한계에 부딪힌다. 효과적인 시각화가 어렵기 때문이다.

Table 4.1에서 잠재변인 간 상관계수를 살펴보면 상호간 상관관계가 뚜렷하게 나타난다 (상관계수 절대값이 0.46과 0.71에 걸쳐있다). 이것은 변수군집을 대표하는 잠재변인들에 근원적 공통요인이 개입되어 있음을 의미한다.

혹시 분석에서 제외되었던 3개 변수가 그런 역할을 하는 것이 아닐까? 이런 가능성을 검토하기 위하여 각 변수군집 별 주성분분석에서 제외되었던 Age, Weight, Height 등을 추가변수로 넣어보았다. Figure 4.4에서 변수군집 행렬도들을 보라. 모든 그래프에서 개별 변수그룹이 Weight 또는 Height와 관련이 있음을 볼 수 있다. 상대적으로 Age와의 관련성은 작게 나타난다.

이에 따라 분석변수들을 Weight, Height, Age에 회귀하여 설명되는 부분을 추출해낸 뒤 남는 잔여분을 분석변수로 재정의하여 잠재적 변수군집화에 투입할 필요가 있다. 안정성 평가 결과 6–8개 변수군집이 적정한 것으로 나타났다. 여기서는 변수군집 수를 6개로 하였다. Figure 4.5에서 덴드로그램을 보라. 변수들의 멤버십이 일부 바뀐 것을 볼 수 있다 (Figure 4.3의 덴드로그램과 비교).

Table 4.2는 6개 잠재변인 간 상관관계를 보여준다. 공통요인이 상당부분 제거되었음을 볼 수 있다 (상관계수 절대값이 0.03과 0.34 사이에 걸쳐있다). 결론적으로 body 데이터의 21개 신체부위의 고유한 잠재적 공간은 6차원으로 되어 있다고 볼 수 있다.

4.3. Wine

ClustOfVar 팩키지의 wine 자료는 21종 포도주에 대한 29개의 수치형 변수(Odor.Intensity.before.shaking, Aroma.quality.before.shaking, . . .)와 2개의 범주형 변수(Label, Soil)로 구성되어 있다. 여

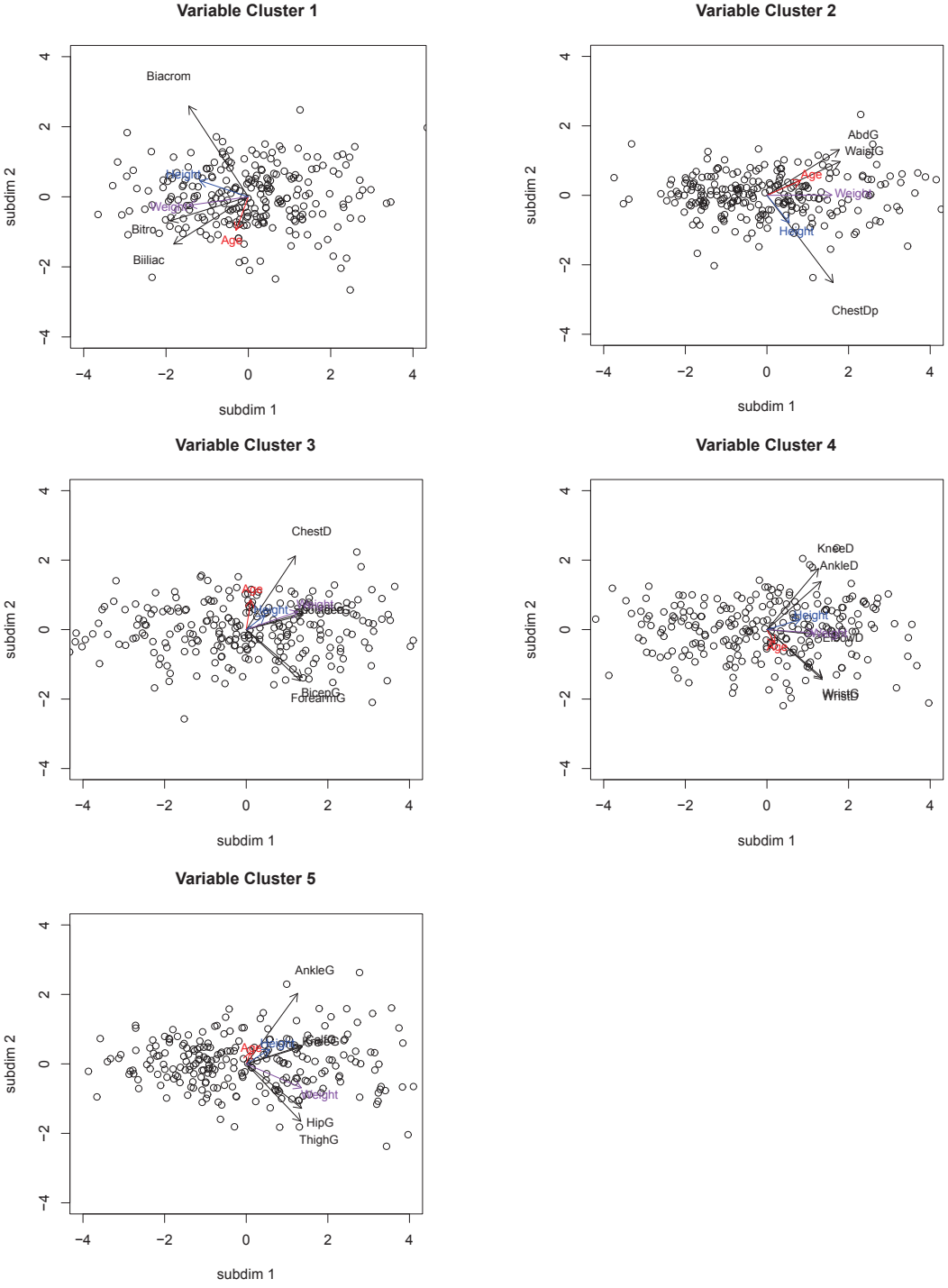

Figure 4.4. Biplots of the body parts data by variable clusters.

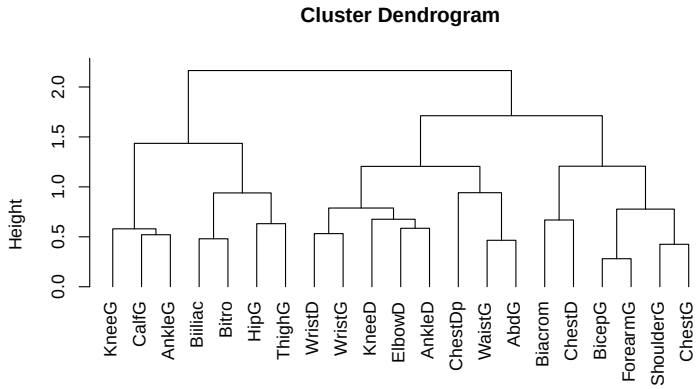

Figure 4.5. Dendrogram for variable clustering of the body parts data after removing the effects of Weight, Height and Age.

Table 4.2. Correlation matrix of the body parts data with residual variables

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster 1	1.00	−0.07	0.12	−0.08	−0.21	−0.03
Cluster 2	−0.07	1.00	0.16	−0.06	−0.25	0.24
Cluster 3	0.12	0.16	1.00	−0.34	−0.21	−0.14
Cluster 4	−0.08	−0.06	−0.34	1.00	0.30	0.32
Cluster 5	−0.21	−0.25	−0.21	0.30	1.00	−0.12
Cluster 6	−0.03	0.24	−0.14	0.32	−0.12	1.00

기서는 29개의 수치형 변수와 범주형 변수 Soil을 분석에 포함하기로 한다. Soil은 포도원의 토양으로 “Reference”, “Env1”, “Env2”, “Env4”의 4개 범주 중에서 값을 취한다. 앞의 사례들에서와 같이 ClustOfVar 팩키지를 써서 변수군집화를 하여 4개 또는 7개의 변수군집이 후보로 나타났다. 여기서 4개를 취하기로 한다.

범주형 변수 Soil은 수치형 변수 Odor.Intensity.before.shaking, Spice.before.shaking, Odor.Intensity, Spice, Bitterness와 함께 제1변수군집에 속하는 것으로 나타났고 4개 범주에 부여된 수치는 −0.032 (Reference), −0.279 (Env1), −0.091 (Env2), 1.316 (Env4)였다. 이것은 “Env4”가 수량화된 Soil (= Soil*)의 플러스 방향이 있고 나머지 범주들이 마이너스 방향에 있음을 의미한다.

Figure 4.6은 변수군집 1에 대한 주성분 행렬도이다 (효과적인 시각적 전달을 위하여 개별변수 및 잠재변인의 화살 길이를 5배로 하였다). 수량화된 범주형 변수 Soil*가 수치형 변수 Bttr (= Bitterness)와 함께 제1잠재변인에 가장 근접하고 있음을 볼 수 있다. 나머지 변수군집들은 모두 수치형 변수들로 구성되어 있으므로 분석결과의 기술을 생략한다.

4.4. 모의생성 자료

앞의 두 사례에서는 잠재변인 간 관계를 시각화하지 않고 상관계수로만 보았다. 이들 계수들이 대체로 0에 가깝기 때문에 선형적으로는 별 의미가 없다. 그렇기에 잠재변인 간 산점도에서 주목할 가치가 남아있을 가능성이 적다. 그러나 그렇지 않은 경우가 있을 수 있다. 다음의 모의생성 자료를 앞의 방법론으로 분석해 보자.

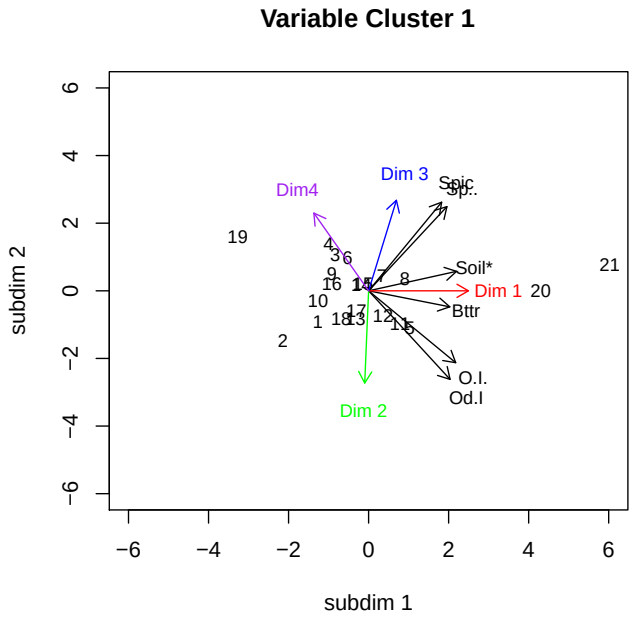

Figure 4.6. Biplot of the wine data: Variable Cluster 1.

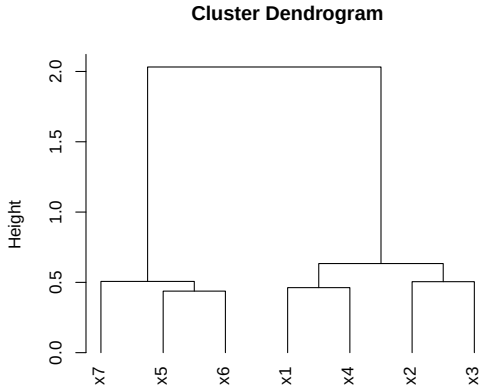

Figure 4.7. Dendrogram for variable clustering of the simulated data.

1) $N(0,1)$ 분포에서 n (=100)개 자료값을 생성하고 이를 제 1 진(眞)잠재변인 S_1^*로 표기한다. 그리고 제2진잠재변인 S_2^*를 $S_1^{*2}-1$로 놓는다. S_1^*와 S_2^*의 평균은 모두 0이고 $\mathrm{Cov}(S_1^*, S_2^*)=0$이다.

$$X_j = \begin{cases} S_1^* + \epsilon_j^*, & j = 1,2,3,4, \\ S_2^* + \epsilon_j^*, & j = 5,6,7, \end{cases}$$

여기서 $\epsilon_1^*, \ldots, \epsilon_7^*$은 길이 n의 $N(0,1)$ 변량이다.

변수군집화 덴드로그램은 Figure 4.7과 같으며 이로부터 7개 변수가 $\{X_1, X_2, X_3, X_4\}$와 $\{X_5, X_6, X_7\}$

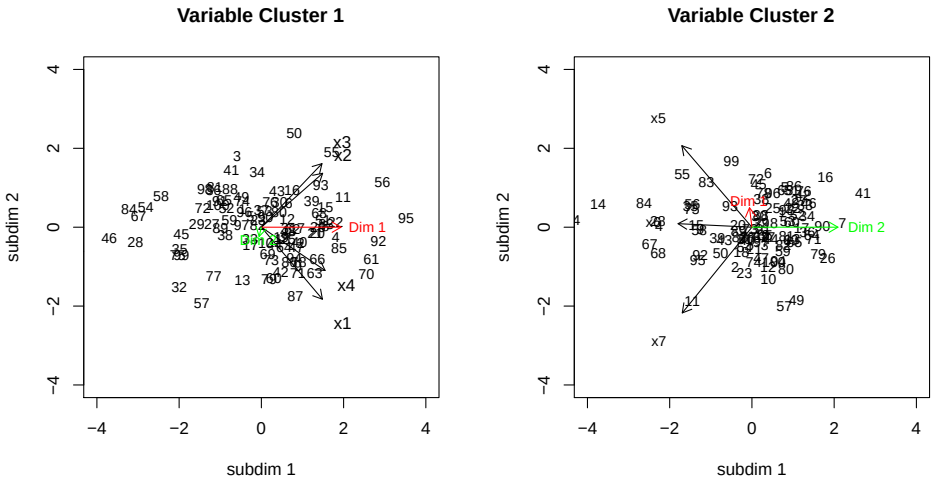

Figure 4.8. Biplots of the simulated data by variable clusters.

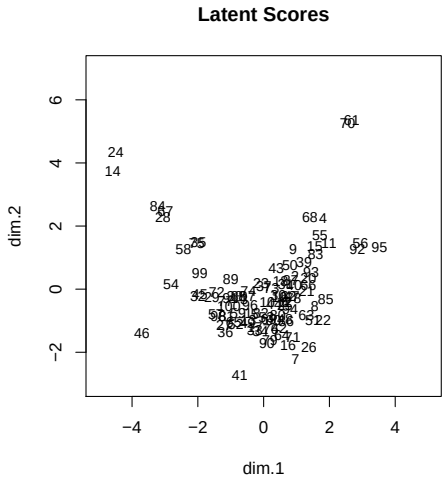

Figure 4.9. Scatterplot of two latent variables from the simulated data.

로 묶이는 것을 볼 수 있다. Figure 4.8은 변수군집 $\{X_1, X_2, X_3, X_4\}$과 $\{X_5, X_6, X_7\}$의 주성분분석 행렬도이다.

Figure 4.9는 $\{X_1, X_2, X_3, X_4\}$의 대표 잠재변인 S_1과 $\{X_5, X_6, X_7\}$의 대표 잠재변수인 S_2 간 산점도이다. 2개의 잠재변인 간 볼록 2차식의 관계가 재현됨을 볼 수 있다. 이 사례가 주는 교훈은 K개(= 잠재군집 수) 잠재변인 간 관계를 살펴볼 필요가 있다는 것이다.

5. 맺음말

이 연구는 변수 수 p가 다변량 자료의 시각화를 위해 1) 변수군집화를 통해 자료세트를 세로 방향으로

분할하고 2) 주성분분석 행렬도로 각 부세트(subset) 자료를 잠재변인으로 축약하여 시각화하며 3) 부세트별 잠재변인들 간 연관성을 파악할 것을 제안하였다.

제안 방법은 비지도학습(unsupervised learning)의 일종이지만, 지도학습에서는 목표변수를 제외한 설명변수들 $\mathbf{x}_1, \ldots, \mathbf{x}_p$의 자료공간을 탐색하는 데 유용할 것으로 기대한다. 활용방식은 1) 부세트별 잠재변인과 목표변수 간 상관성에 기초하여 소수의 잠재변인을 선별하고 나서 2) 앞 단계에서 선별된 잠재변인들을 지도학습모형에 입력변수로 투입하는 것이다. 이로써 형태적으로는 은닉 층(hidden layer)이 1개인 신경망 모형과 유사한 모습이 된다. 그러나 제안 방법은 신경망과는 달리 은닉노드를 탐색적 자료분석(exploratory data analysis)으로 구성해낸다는 점에서 나름의 특성이 있다.

References

Benzecri, J. P. (1992). *Correspondence Analysis Handbook*, Marcel Dekker, New York.

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012). ClustOfVar: an R package for the clustering of variables, *Journal of Statistical Software*, **50**, 1–16.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, J. (2013). Package 'ClustOfVar'. R Foundation for Statistical Computing, URL https://cran.r-project.org/mirrors.html.

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

Vigneau, E. and Chen, M. (2015). Package 'ClustVarLV'. R Foundation for Statistical Computing, from: https://cran.r-project.org/mirrors.html.

Vigneau, E., Chen, M., and Qannari, E. M. (2015). ClustVarLV: an R package for the clustering of variables around latent variables, *The R Journal*, **7**, 134–148.

Vigneau, E. and Quannari, E. M. (2003). Clustering of variables around latent components, *Communications in Statistics – Simulation and Computation*, **32**, 1131–1150.

분석변수들의 잠재공간 표현

허명회[a,1]

[a]고려대학교 통계학과

(2017년 5월 8일 접수, 2017년 6월 10일 수정, 2017년 6월 10일 채택)

요 약

다변량 자료에서 변수 수 p가 큰 경우 주성분분석 등 통상적인 차원축소는 효과적이지 못할 수 있다. 효과적인 시각화가 되려면 축소공간의 차원이 2–3 정도이어야 하는데, 관측개체의 잠재적 차원이 이보다 훨씬 큰 경우가 있기 때문이다. 이 논문은 분석변수들을 다수의 잠재 차원에 분할하여 차원축소적 방법으로 탐색하고 부분들의 유기적 관계를 시각화하는 이단계 작업을 제안한다. 분석변수들을 잠재 차원에 분할하는 "잠재변인 변수군집화" 방법으로는 R 팩키지 ClustOfVar를 쓰고 개별 변수군집의 시각화를 위해서 주성분분석 행렬도(biplot)를, 개별 변수군집과 외부 잠재변인 또는 외적 변수 간 관계의 시각화를 위해서는 추가변수 끼워넣기(embedding supplementary variables) 기법을 활용한다.

주요용어: 데이터 시각화, 변수군집화, 잠재변인, 주성분분석, 행렬도, 추가변수

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