가상점을 활용한 다차원척도법의 재해석

Eun-seong Kim[1], Yong-Seok Choi[2]

1. Master's course, Department of Statistics, Pusan National University
2. Professor, Department of Statistics, Pusan National University

1. Introduction

❖ 다차원척도법(multidimensional scaling)이란 다차원 공간의 개체간의 거리를 나타내는 자료로부터 그들의 유사성 또는 비유사성을 저차원 공간에 기하적으로 나타내어 그들의 관계를 탐색적으로 살펴보는 다변량 그래프적 기법이다. 다차원척도법에서는 저차원 공간을 형상공간(configuration space)이라 하며 여기에 개체를 기하적으로 나타낸 것을 다차원척도법도(multidimensional scaling map)라 한다.

❖ 일반적으로 계량형 다차원척도법은 양적자료(quantative data)에 적용하는데 이진수 자료(binary data)는 질적자료(qualitative data)임에도 불구하고 계량형 다차원척도법을 적용할 수 있다. 이는 이진수 자료에서 비유사성은 유사성 측도를 위해 보편적으로 이용하는 단순매칭계수로부터 계산하고 이를 두 개체간의 거리 개념에서 보면 이진수 자료의 제곱유클리드거리를 전체 항목 수로 나눈 것과 동일하기 때문이다.

❖ 다차원척도법을 통해서는 각각의 개체들에 대한 군집화만을 파악할 수 있는데 이러한 군집들의 특징을 파악하기 위해서는 가상점(pseudo-sample)을 활용한 개별 변수의 추가적인 표현이 요구된다.

❖ 다차원척도법 공간상에 가상점을 표현하는 기존의 방법인 Gower와 Hand(1996)가 제안한 대체법은 개별 변수에 해당하는 관측값만을 τ(개별 변수가 가질 수 있는 값)로 대체하여 가상점의 중심을 계산하고 이를 다차원척도법 좌표공간에 투영(projection)하여 가상점을 표현한다.

❖ 이진수 자료는 가상점의 중심이 1의 비율에 따라 결정되므로, 다차원척도법 공간상에 가상점을 표현하기 위해서는 각각의 변수에 해당하는 관측값이 0인 경우와 1인 경우를 분할한 행렬 각각을 가상점이라 두고 중심을 계산하는 분할법을 이용해야 한다.

2. Multidimensional scaling

❖ p개의 변수로 n개의 개체에 대해 얻은 이진수 자료 $\mathbf{X}=(x_{ik}), i=1,\cdots,n; k=1,\cdots,p$ 는 다음과 같다.

$$x_{ik}=\begin{cases}1: i\text{번째 개체가 } k\text{번째 변수의 성질을 만족하는 경우}\\ 0: \text{그 외의 경우}\end{cases}$$

❖ 이진수 자료에서 다차원척도법을 위한 비유사성은 두 개체간의 유사성을 나타내는 단순매칭계수를 이용하여 계산하는데 이는 이진수 자료의 제곱유클리드거리를 p로 나눈것과 일치한다. 따라서 이진수 자료에서 제곱유클리드거리를 비유사성으로 하므로 계량형 다차원척도법을 실시할 수 있다.

❖ 이진수 자료 $\mathbf{X}$의 i번재 개체 개체 $\mathbf{x}_i=(x_{i1},\cdots,x_{ip})^t$ 와 j번째 개체 $\mathbf{x}_j=(x_{j1},\cdots,x_{jp})^t$ 간의 제곱유클리드거리는 다음과 같다.

$$d_{ij}^2=(\mathbf{x}_i-\mathbf{x}_j)^t(\mathbf{x}_i-\mathbf{x}_j),\ i,j=1,\cdots,n$$

❖ 계량형 다차원척도법의 대표적인 알고리즘인 토거선 알고리즘(Torgerson, 1958)에 대해 정리하자.

[1단계] 비유사성 행렬로 $\mathbf{D}=(d_{ij}^2),\ i,j=1,\cdots,n$ 부터 행렬 $\mathbf{A}=(a_{ij})$을 계산한다. 행렬 $\mathbf{A}$의 원소는 $a_{ij}=-\frac{1}{2}d_{ij}^2$ 이다.

[2단계] 행렬 $\mathbf{A}$로 부터 이중-중심화 행렬은 $\mathbf{B}=(b_{ij})=\mathbf{HAH}$와 같이 나타나게 된다.

여기서 $b_{ij}=a_{ij}-\bar{a}_{i\cdot}-\bar{a}_{\cdot j}+\bar{a}_{\cdot\cdot}$ 이며, $\bar{a}_{i\cdot}=\sum_{j=1}^{n}x_{ij}/\mathrm{n}$ 은 행렬 $\mathbf{A}$의 i번째 행의 평균이고, $\bar{a}_{\cdot j}=\sum_{i=1}^{n}x_{ij}/\mathrm{n}$ 은 각각 행렬 $\mathbf{A}$의 j번째 열의 평균이고, $\bar{a}_{\cdot\cdot}=\sum_{i=1}^{n}\sum_{j=1}^{n}x_{ij}/n^2$은 $\mathbf{A}$의 모든 원소의 평균이다.

$\mathbf{H}=\mathbf{I}_n-n^{-1}\mathbf{1}_n\mathbf{1}_n^t$ 이며 $\mathbf{I}_n$은 단위행렬이고 , $\mathbf{1}_n$은 모든 원소가 1인 $n\times 1$ 벡터이다.

[3단계] 차원 축소된 형상 공간의 좌표를 얻기 위해서는 이중-중심화 행렬 B의 스펙트럼 분해 $\mathbf{B}=\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^t$를 계산한다. 여기서 $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$는 $\lambda_1\geq\cdots\geq\lambda_n$ 의 관계를 갖는 고유값을 대각원소로 하는 대각행렬이며, $\mathbf{V}$는 크기 $n\times n$ 의 직교행렬이다.

[4단계] 행렬 $\mathbf{B}$의 스펙트럼분해로부터 $\mathbf{s}(\leq \mathbf{p})$ 개의 고유값과 이에 대응하는 고유벡터를 가지고 크기가 $n\times s$ 인 s차원의 형상공간의 좌표는 $\mathbf{C}_{(s)}=\mathbf{V}_{(s)}\boldsymbol{\Lambda}_{(s)}^{1/2}$ 와 같이 나타나고, $\mathbf{V}_{(s)}$는 $n\times s$의 행렬이고, $\boldsymbol{\Lambda}_{(s)}^{1/2}=\mathrm{diag}(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_s})$ 이다.

❖ 토거선의 알고리즘을 통해 차원 축소된 s차원 다차원척도법의 근사적합도는 다음과 같다.

$$\frac{\sum_{r=1}^{s}\lambda_r}{\sum_{r=1}^{p}\lambda_r}\times\mathbf{100}(\%)$$

근사적합도를 다차원척도법의 설명력이라 하고, 값이 70% 이상이면 원자료를 잘 설명한다고 할 수 있다.

3. Representation of pseudo-sample

3.1 Gower와 Hand(1996)가 제안한 대체법에 의한 가상점 표현

❖ p개의 연속형 변수로 n개의 개체에 대해 얻은 자료행렬을 $\mathbf{X}=(x_{ik}),\ i=1,\cdots,n; k=1,\cdots,p$ 라 하자.

이때 k번째 변수에 대한 정보를 얻기 위해서는 주어진 자료 행렬의 k번째 변수에 해당하는 관측값만을 모두 τ(k번째 변수가 가질 수 있는 값)로 대체한 n개의 가상점 $\mathbf{X}^*=(x_{ik}^*)$을 다음과 같이 고려해야 한다.

$$\mathbf{X}^*=\begin{pmatrix}x_{11}^*, & \cdots, & x_{i;k-1}^*, & \tau, & x_{1;k+1}^*, & \cdots, & x_{ik}^*\\ & & & \vdots & & & \\ x_{i1}^*, & \cdots, & x_{i;k-1}^*, & \tau, & x_{i;k+1}^*, & \cdots, & x_{ik}^*\\ & & & \vdots & & & \\ x_{n1}^*, & \cdots, & x_{n;k-1}^*, & \tau, & x_{n;k+1}^*, & \cdots, & x_{ik}^*\end{pmatrix}$$

❖ 자료행렬 $\mathbf{X}$의 k번째 변수에 해당하는 관측값만을 모두 τ로 대체한 n개의 가상점의 중심은 다음과 같다.

$$\bar{\mathbf{x}}^*=(\bar{x}_1^*,\cdots\bar{x}_{k-1}^*,\tau,\bar{x}_{k+1}^*,\cdots\bar{x}_p^*)^t$$

여기서 $\bar{x}_k^*=\sum_{i=1}^{n}x_{ij}^*/\mathrm{n}, k=1,\cdots,p$ 는 가상점 $\mathbf{X}^*$의 k번째 열의 평균이므로 τ이다.

❖ n개 가상점의 중심과 기존의 n개 좌표점과의 제곱거리를 나타내는 $n\times 1$ 벡터 $\mathbf{a}=(a_1,\ \cdots,\ a_n)^t$ 의 원소는 다음과 같다.

$$a_i=(\bar{\mathbf{x}}_i^*-\bar{\mathbf{x}}^*)^t(\bar{\mathbf{x}}_i^*-\bar{\mathbf{x}}^*),\ i=1,\cdots,n$$

❖ 가상점의 중심을 s차원으로 축소된 다차원척도법 좌표공간에 투영한 $s\times 1$ 좌표벡터는 다음과 같다.

$$\mathbf{c}_k=\boldsymbol{\Lambda}_{(s)}^{-1}\mathbf{C}_{(s)}^t\left[-\frac{1}{2}\mathbf{a}-\frac{1}{n}\mathbf{A}\mathbf{1}_n\right]$$

3.2 분할법에 의한 가상점 표현

❖ 여러 개의 관측값이 아닌 0 과1의 두 개의 관측값만을 갖는 이진수 자료의 경우 k번째 변수에 해당하는 관측값을 모두 τ(0또는1)로 대체하여 가상점의 중심을 구하는 대체법이 아닌 k번째 변수에 해당하는 관측값이 0인 경우와 1인 경우를 분할한 행렬 각각을 가상점이라 두고 중심을 구하는 분할법을 이용해야 한다. 이는 이진수 자료에서는 가상점의 중심이 1의 비율에 따라 결정되므로 0과 1의 분할된 가상점의 열평균을 계산해야한다.

❖ 분할법 알고리즘에 대해 간략히 정리하자.

[1단계] 이진수 자료 X가 주어졌을 때 k번째 변수의 분할계획행렬 $\mathbf{G}_k$ 를 계산한다.

분할계획행렬 $\mathbf{G}_k=(g_{ih}^{(k)}),\ i=1,\cdots,n,\ k=1,\cdots,p,h=1,2$ 의 각 원소는 다음과 같다.

$$g_{ih}^{(k)}=\begin{cases}1: i\text{번째 개체가 } k\text{번째 변수에대해 } h\text{번째 범주에 해당하는 경우}\\ 0: \text{그 외의 경우}\end{cases}$$

[2단계] k번째 변수의 관측값이 τ(0 또는 1) 인 개체 수와 열별합을 계산한다.

$$\mathbf{D}_k=\mathbf{G}_k^t\mathbf{G}_k=\begin{bmatrix}\sum_{i=1}^{n}g_{i1}{}^{(k)} & 0\\ 0 & \sum_{i=1}^{n}g_{i2}{}^{(k)}\end{bmatrix}$$

여기서 $\sum_{i=1}^{n}g_{i1}{}^{(k)}$은 이진수 자료 $\mathbf{X}$의 k번째 변수의 관측값이 1인 개체수이고, $\sum_{i=1}^{n}g_{i2}{}^{(k)}$는 0인 개체수이다.

k번째 변수의 관측값이 0인 경우와 1인 경우를 분할한 가상점의 열별합은 $\mathbf{G}_k^t\mathbf{X}$ 이다.

[3단계] k번째 변수의 값이 τ(0 또는 1) 인 분할된 가상점의 중심을 계산한다.

$$\mathbf{Y}_k=\mathbf{D}_k^{-1}\ \mathbf{G}_k^t\mathbf{X}=[\mathbf{y}_1,\mathbf{y}_2]^t$$

여기서 $\mathbf{y}_1$은 k번째 변수의 관측값이 1인 경우 열평균을 나타내는 $p\times 1$벡터이고, $\mathbf{y}_2$는 0인 경우 열평균을 나타내는 $p\times 1$ 벡터이다.

[4단계] k번째 변수의 관측값이 τ(0 또는 1)인 분할된 가상점의 중심과 기존의 n개 좌표점과의 제곱거리벡터는 $n\times 1$ 벡터 $\mathbf{a}_r=(a_{1(r)},\cdots,a_{n(r)})^t$의 원소는 다음과 같다.

$$\boldsymbol{a}_{i(r)}=(\mathbf{x}_i-\boldsymbol{y}_r)^t(\mathbf{x}_i-\boldsymbol{y}_r),\ i=1,\cdots,n,\ r=1,2$$

여기서 $\mathbf{a}_1$은 k번째 변수의 관측값이 1인 가상점의 중심과 기존의 n개 좌표점과의 제곱거리벡터이고, $\mathbf{a}_2$는 k번째 변수의 관측값이 0일때의 제곱거리벡터이다.

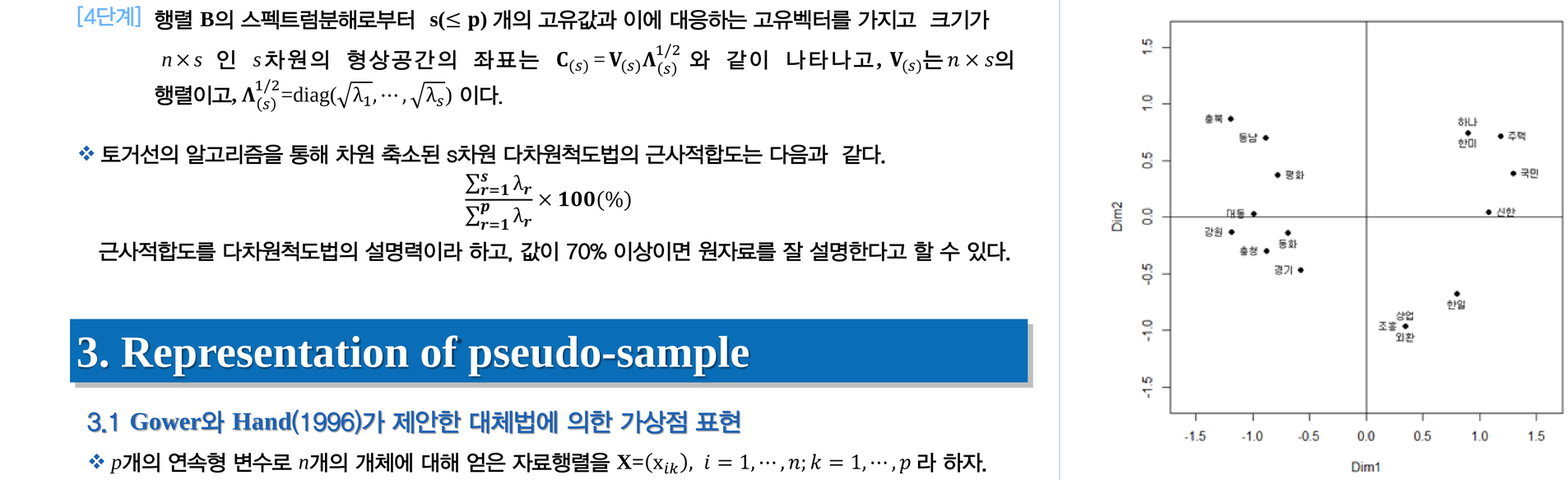
[5단계] 가상점의 중심을 s차원으로 축소된 다차원척도법 좌표공간에 투영한 $s\times 1$좌표벡터는 다음과 같다.

$$\mathbf{c}_r=\boldsymbol{\Lambda}_{(s)}^{-1}\mathbf{C}_{(s)}^t\left[-\frac{1}{2}\mathbf{a}_r-\frac{1}{n}\mathbf{A}\mathbf{1}_n\right],\ r=1,2$$

[6단계] 변수 개수 p개 만큼 [1단계]–[5단계]를 반복한다.

4. Illustrations

❖ 중앙일보(1988. 6. 29)에 실린 〈살아난 은행들도 조마조마〉라는 제목의 퇴출된 은행 및 시중 17개 은행의 경영평가 결과에 대해 8가지 항목으로 가공된 이진수 자료의 2차원 계량형 다차원척도법도는 〈그림 1〉이며, 가상점을 추가한 계량형 다차원척도법도는 〈그림 2〉이다. 이에 대한 가상점의 좌표벡터는 〈표 1〉에 나타내었다.



〈그림 1〉 계량형 다차원척도법도

〈그림 2〉 가상점을 추가한 계량형 다차원척도법도

〈표 1〉가상점 좌표벡터

	$\tau=1$		$\tau=0$	
	Dim1	Dim2	Dim1	Dim2
BIS ratio	1.070	0.527	−0.446	−0.220
asset	0.798	−0.104	−0.898	0.117
NPL	0.770	−0.347	−0.539	0.243
NPL ratio	1.024	0.327	−0.559	−0.178
IO	0.202	0.648	−0.141	−0.453
branch	0.319	−0.088	−1.037	0.285
workforce	0.319	−0.088	−1.037	0.285
grade	0.136	−0.202	−0.249	0.370