

연세대학교 통계 데이터 사이언스 학회 ESC 23-2 FALL WEEK6

Multidimensional Scaling & Correspondence Analysis

[ESC 정규세션 학술부] 이상윤 김채성



목차

Part1

1. MDS
2. metric MDS
3. nonmetric MDS

Part2

4. Correspondence Analysis
5. Chi-square Decomposition
6. Interpreting with Biplots





1. MDS

MDS

MDS 란?

차원축소 기법중 하나로, 각 observation의 거리정보를 보존하면서 차원을 축소하는 것이 목적이다. 특히 Euclidean coordinate을 갖지 않는 데이터에 대해 (ranking 같은), 해당 데이터의 거리정보를 유지하며 2/3차원 유클리드 공간에 나타냄으로써 observation간 거리, 즉 유사도를 한눈에 쉽게 확인할 수 있도록 한다.

-> 시각화에 유용!!



MDS

MDS vs PCA

MDS	PCA
<ul style="list-style-type: none">- 임의의 차원 p에서의 거리를 나타내는 Distance matrix / Dissimilarity matrix $D (n \times n)$ 에서 시작-목적: 거리정보를 보존하는 좌표계 찾기- p 차원의 좌표값 출력 ($n \times p$)	<ul style="list-style-type: none">-p 개의 변수로 이루어진 data matrix $X (n \times p)$ 에서 시작-목적: 분산을 보존하는 변수들의 선형결합 찾기-고윳값에 따른 Principle columns

* metric MDS의 경우 PCA와 거의 유사하지만, Nonmetric MDS의 경우 거리의 값이 아닌 ranking 만을 사용한다는 점에서 PCA와 차이가 있음.



MDS

Why MDS?

- distance matrix 에서 출발하므로 어떠한 종류의 distance 를 쓰던 적용 가능 -> 광범위한 분야에서 가능

ex) Euclidean distance, Manhattan distance, log-fold changes(bio), Hamming distance, Great circle distance



MDS

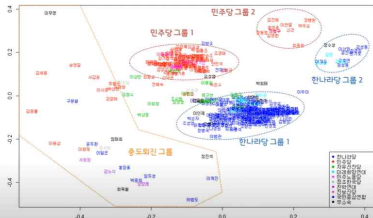
MDS의 사용분야

1. 마케팅

- 제품의 포지셔닝, 소비자 선호도 분석, 시장 세분화 및 표적시장 선정, 신제품 개발, 가격 결정, 광고 연구 영역
 - 그러나 이러한 유용성에 비해 실무에 적용사례 적음. 일부 모델만이 시중에 있으며 컨설팅 회사, 리서치회사에선 좋은 MDS를 제공하는 곳이 거의 없음.
- (출처:이화여대 경영학과 김영찬 교수)

2.

MDS Example



강형선, 박영준, 조수근, 김성범, (2013), 대한민국 18대 국회의원 의정활동 분석, 한국경영과학회 추계학술대회





2. Metric MDS

Metric MDS

Metric MDS의 개념

p-dimension에서 n개의 점들사이의 거리가 기존 distance matrix D 의 원소와 최대한 비슷하도록 하는 좌표를 찾는 것

-> D 가 Euclidean distance를 나타낸 distance matrix인 경우, MDS=PCA

+ D 에서 original coordinate을 찾기 (recovery)



Metric MDS

Algorithm

1. start with distances d_{ij}
2. define $\mathcal{A} = -\frac{1}{2}d_{ij}^2$
3. put $\mathcal{B} = (a_{ij} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet})$
4. find the eigenvalues $\lambda_1, \dots, \lambda_p$ and the associated eigenvectors $\gamma_1, \dots, \gamma_p$ where the eigenvectors are normalized so that $\gamma_i^T \gamma_i = 1$.
5. Choose an appropriate number of dimensions p (ideally $p = 2$)
6. The coordinates of the n points in the Euclidean space are given by $x_{ij} = \gamma_{ij} \lambda_j^{1/2}$ for $i = 1, \dots, n$ and $j = 1, \dots, p$.



Metric MDS

Algorithm

1. distance matrix D 구하기(dij)

$$d_{ij}^2 = (x_i - x_j)^\top (x_i - x_j) \quad \text{OR} \quad d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2.$$

2. A 구하기 (aij)

$$a_{ij} = -\frac{1}{2}d_{ij}^2$$



Metric MDS

Algorithm

3. distance matrix D와 B의 관계식 구하기

$$B = HAH$$

pf) ① $d_{ij}^2 = x_i^\top x_i + x_j^\top x_j - 2x_i^\top x_j$
 $= b_{ii} + b_{jj} - 2b_{ij}.$

(B를 original coordinate(x)로 나타낸 식)

$$b_{ij} = \sum_{k=1}^p x_{ik} x_{jk} = x_i^\top x_j$$

② $\frac{1}{n} \sum_{i=1}^n d_{ij}^2 = \frac{1}{n} \sum_{i=1}^n b_{ii} + b_{jj}$

③ $\frac{1}{n} \sum_{j=1}^n d_{ij}^2 = b_{ii} + \frac{1}{n} \sum_{j=1}^n b_{jj}$

④ $\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2 = \frac{2}{n} \sum_{i=1}^n b_{ii}.$

① - ② - ③ + ④

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2).$$

$$= a_{ij} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet}$$

$$\Leftrightarrow B = HAH$$



Metric MDS

Algorithm

4. $HAH=B$ 로 구한 B 를 SVD

$$B = \Gamma \Lambda \Gamma^T$$

5. $B=XX'$ 를 이용하여 X 구하기

$$B=XX^T=\Gamma\Lambda\Gamma^T \rightarrow X=\Gamma\Lambda^{\frac{1}{2}} \quad (\text{original coordinate. } n \times p)$$

$$\text{rank}(B)=\text{rank}(XX^T)=\text{rank}(X)=p$$

$n-p$ 개의 고윳값은 0



Metric MDS

적절한 p 값 (optimal dimension)?

Proportion of variation explained by p-dimension 에 따라 결정

$$\frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^{n-1} \lambda_i}$$

(B가 Positive semidefinite 가 아닌 경우)

$$\frac{\sum_{i=1}^p \lambda_i}{\sum (\text{"positive eigenvalues"})}$$



Metric MDS

D 가 symmetric matrix 일때 (when D is not distance matrix)

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{\frac{1}{2}}$$

i.e B=HCH

이렇게 놓고 앞과 동일한 과정 거치면 됨.



Metric MDS

Metric MDS=PCA (projection 으로서의 MDS)

\mathcal{X}_1 : 기존 p-dimension에서 제시된 distance matrix D를 갖는 coordinate matrix χ 를 더 낮은 차원인 k-dimension에 representation한 것
이때 어떤 orthogonal matrix L에 대하여 $L_{p \times p} = (L1_{p \times k} | L2_{p \times (p-k)})$ 이라 하면

$$\mathcal{X}_1 = \mathcal{X}L_1$$

이는 χ 를 L_1 의 column space로 projection 한 것으로 볼 수 있다. 이것의 distance matrix인 D_1 과 기존 D의 차이는 다음의 값으로 측정할 수 있다.

$$\phi = \sum_{i,j=1}^n (d_{ij} - d_{ij}^{(1)})^2. \quad \mathcal{D}_1 = (d_{ij}^{(1)})$$

이때 이 phi 값이 최소가 되는 k는 χ_1 이 χ 의 first k principal factors일 때!!





3. Nonmetric MDS

Nonmetric MDS

Nonmetric MDS 의 개념

거리의 rank order 을 사용하는 MDS로 iterative process 를 통해 좌표를 구한다.

Metric MDS 가 거리정보를 유지한 채 차원을 축소시키는 것이었다면, Nonmetric MDS는 거리의 **'ranking'**을 유지한 채 차원을 축소시키는 것!

따라서 비언어적데이터, 이미지 등 Euclidean distance가 아닌 경우에도 사용이 가능하며, 추가적인 scaling이나 monotonic transformation을 적용해도 결과가 동일하다.



Nonmetric MDS

Notation

$$d_{ij} = f(\delta_{ij})$$

monotonic increasing function

distance

dissimilarity

χ_k : k th coordinate configuration(configuration).
k=0 의 경우는 initial configuration

$d_{ij}^{(k)}$: kth coordinate coordinate에서 구한 point i 와 j 사이의 거리



Nonmetric MDS

Algorithm : Shepard-Kruskal Algorithm

1. Choose an initial configuration. \mathcal{X}_0
2. Find d_{ij} from the configuration. (dij 계산 by f)
3. Fit \hat{d}_{ij} , the disparities, by the PAV algorithm. (f가 monotone 이 되도록 수정)
4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent.
5. Go to 2. (update configuration)

4.5. Evaluate changes of the iteration
(STRESS)



Nonmetric MDS

Algorithm

1. Choose an initial configuration

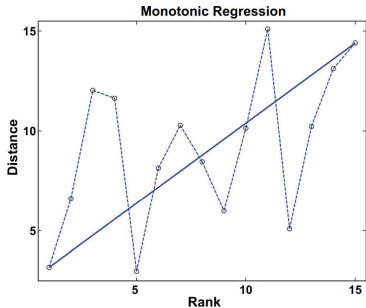
임의로 고르면 됨. metric MDS의 결과를 써도 됨

2. Find d_{ij} from the configuration.

$$d_{ij} = f(\delta_{ij})$$

(f 는 monotonic)

But, 이 경우 f 는 monotonic하지 않음 -> 수정 필요!!



Nonmetric MDS

3. PAV algorithm(pool-adjacent violators)

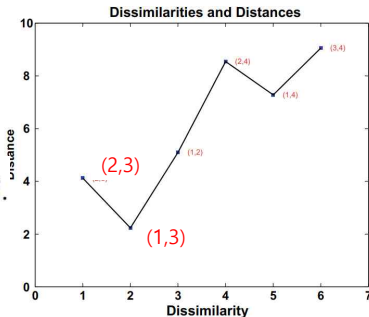
monotonic 하지 않은 경우, weak monotonicity라도 성립하도록 수정하는 algorithm.

monotonicity가 어긋난 점과 그 바로 앞의 점 두 개의 distance를 모두 둘의 평균값으로 수정. 이렇게 수정된 값을 \hat{d} 을 씌움 -> disparity라 부름

ex)

$$\hat{d}_{13} = \hat{d}_{23} = \frac{d_{13} + d_{23}}{2} = \frac{2.2 + 4.1}{2} = 3.17.$$

※ 이는 point 3의 좌표를 바꾸는 것과 같은데, 이에 따라 d_{34} 의 값도 바뀐다. 이것이 monotonicity를 해치지 않는지 체크해야한다.



Nonmetric MDS

4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent/Newton-Raphson procedure

$$x_{il}^{NEW} = x_{il} + \alpha \left(1 - \frac{\hat{d}_{ij}}{d_{ij}} \right) (x_{jl} - x_{il}), \quad l = 1, \dots, p^*. \quad (\text{point } i \text{를 point } j \text{로 update 하는 경우})$$

$$x_{il}^{NEW} = x_{il} + \frac{\alpha}{n-1} \sum_{j=1, j \neq i}^n \left(1 - \frac{\hat{d}_{ij}}{d_{ij}} \right) (x_{jl} - x_{il}), \quad l = 1, \dots, p^*. \quad (\text{point } i \text{를 다른 모든 point로 update 하는 경우})$$

α 는 step width로 보통 0.2를 사용.



Nonmetric MDS

4.5. Evaluate changes of the iteration (STRESS)

STRESS : 새롭게 얻은 configuration과 given dissimilarities의 차이를 나타내는 지표. 두 가지 버전 있음.

$$STRESS1 = \left(\frac{\sum_{i < j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i < j} d_{ij}^2} \right)^{\frac{1}{2}} \quad STRESS2 = \left(\frac{\sum_{i < j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i < j} (d_{ij} - \bar{d})^2} \right)^{\frac{1}{2}}$$

판단 기준: 어느 정도 작으면 멈춰도 됨. 혹은 아래 기준 참고 (S means STRESS)

$S > 20\%$, poor; $S = 10\%$, fair; $S < 5\%$, good; $S = 0$, perfect.

* decision of appropriate dimension p *

p에 따른 minimum STRESS 를 plot으로 그린 뒤, 'elbow' 가 생기는 지점의 p를 고른다.



Nonmetric MDS

정리

- distance(dissimilarity)의 ranking을 유지하며 차원을 축소시키려 함(축소된 차원의 좌표를 구하려 함)
- monotonicity가 violated된 부분을 조정하면서(PAV) + STRESS를 최소화하는 configuration을 찾는 것이 목표.
- 이를 위해 iterative procedure를 사용하여 configuration을 계속 update 하고, STRESS가 충분히 작다면 iteration을 종료. 해당 좌표계로 나타냄.



Nonmetric MDS



Example - car marks

Table 17.3 Dissimilarities δ_{ij} for car marks

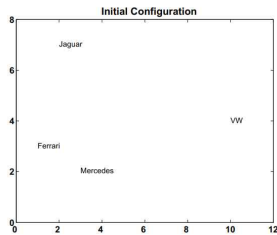
	j	1	2	3	4
i		Mercedes	Jaguar	Ferrari	VW
1	Mercedes	—			
2	Jaguar	3	—		
3	Ferrari	2	1	—	
4	VW	5	4	6	—

→ find monotonically increasing f
such that $d_{ij} = f(\delta_{ij})$

Table 17.4 Initial coordinates for MDS

i		x_{i1}	x_{i2}
1	Mercedes	3	2
2	Jaguar	2	7
3	Ferrari	1	3
4	VW	10	4

$:\mathcal{X}_0$



1. Choose an initial configuration.
2. Find d_{ij} from the configuration.
3. Fit \hat{d}_{ij} , the disparities, by the PAV algorithm.
4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent.
5. Go to 2.

Nonmetric MDS

Example - car marks

1. Choose an initial configuration.
2. Find d_{ij} from the configuration.
3. Fit \hat{d}_{ij} , the disparities, by the PAV algorithm.
4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent.
5. Go to 2.

Table 17.5 Ranks and distances

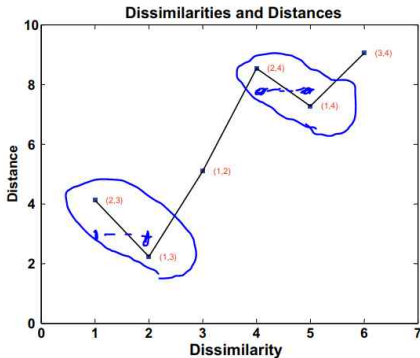
i, j	d_{ij}	$rank(d_{ij})$	δ_{ij}
1, 2	5.1	3	3
1, 3	2.2	1	2
1, 4	7.3	4	5
2, 3	4.1	2	1
2, 4	8.5	5	4
3, 4	9.1	6	6



Nonmetric MDS



Example - car marks



1. Choose an initial configuration.
2. Find d_{ij} from the configuration.
3. Fit \hat{d}_{ij} , the disparities, by the PAV algorithm.
4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent.
5. Go to 2.

$$\hat{d}_{13} = \hat{d}_{23} = \frac{d_{13} + d_{23}}{2} = \frac{2.2 + 4.1}{2} = 3.15$$

Nonmetric MDS

Example - car marks

$$x_{31} = 1 \text{ and } x_{32} = 3.$$

Applying (17.24) yields (for $\alpha = 3$):

$$\begin{aligned} x_{31}^{NEW} &= 1 + \frac{3}{4-1} \sum_{j=1, j \neq 3}^4 \left(1 - \frac{\hat{d}_{3j}}{d_{3j}}\right) (x_{j1} - 1) \\ &= 1 + \left(1 - \frac{3.15}{2.2}\right) (3 - 1) + \left(1 - \frac{3.15}{4.1}\right) (2 - 1) + \left(1 - \frac{9.1}{9.1}\right) (10 - 1) \\ &= 1 - 0.86 + 0.23 + 0 \\ &= 0.37. \end{aligned}$$

Similarly we obtain $x_{32}^{NEW} = 4.36$.

1. Choose an initial configuration.
2. Find d_{ij} from the configuration.
3. Fit \hat{d}_{ij} , the disparities, by the PAV algorithm.
4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent.
5. Go to 2.



Nonmetric MDS

Example - car marks

Table 17.6 STRESS calculations for car marks example

(i, j)	δ_{ij}	d_{ij}	\hat{d}_{ij}	$(d_{ij} - \hat{d}_{ij})^2$	d_{ij}^2	$(d_{ij} - \bar{d})^2$
(2, 3)	1	4.1	3.15	0.9	16.8	3.8
(1, 3)	2	2.2	3.15	0.9	4.8	14.8
(1, 2)	3	5.1	5.1	0	26.0	0.9
(2, 4)	4	8.5	7.9	0.4	72.3	6.0
(1, 4)	5	7.3	7.9	0.4	53.3	1.6
(3, 4)	6	9.1	9.1	0	82.8	9.3
Σ		36.3		2.6	256.0	36.4

Algorithm : Shepard-Kruskal Algorithm

1. Choose an initial configuration. \mathcal{X}_0
2. Find d_{ij} from the configuration. (\hat{d}_{ij} 계산 by f)
3. Fit \hat{d}_{ij} , the disparities, by the PAV algorithm. (f^2 가 monotone 이 되도록 수정)
4. Find a new configuration \mathcal{X}_{n+1} by using the steepest descent.
5. Go to 2. (update configuration)

4.5. Evaluate changes of the iteration (STRESS)



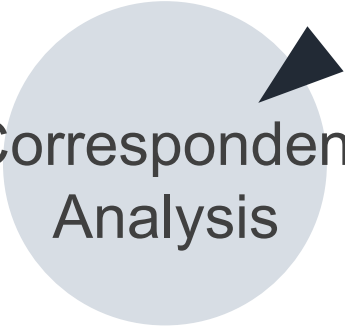
$$STRESS1 = \sqrt{2.6/256} = 0.1$$

$$STRESS2 = \sqrt{2.6/36.4} = 0.27.$$

$S > 20\%$, poor; $S = 10\%$, fair; $S < 5\%$, good; $S = 0$, perfect.

\therefore iteration 조금 더 반복해야함!!





4. Correspondence Analysis

Contingency table

Contingency table (분할표)

범주형 변수인 두 변수에 대해 도수분포표를 2차원으로 확장한 형태의 표.

각 셀은 observed joint frequency를 가진다.

	Male	Female	
Smoke	20	10	30
Non smoke	30	40	70
	50	50	100

Table 12.8 Frequencies of Types of Pottery					
Site	Type				Total
	A	B	C	D	
P0	30	10	10	39	89
P1	53	4	16	2	75
P2	73	1	41	1	116
P3	20	6	1	4	31
P4	46	36	37	13	132
P5	45	6	59	10	120
P6	16	28	169	5	218
Total	283	91	333	74	781

cell probability

marginal probability

conditional probability



Contingency table

카이제곱 검정

분할표의 정보를 활용해서, 범주형 변수에 대해 3가지의 카이제곱 검정을 할 수 있다.

1. 적합도 검정
 - 관측값과 기댓값이 동일한지 검정.
2. 독립성 검정
 - 두 변수가 서로 독립인지 검정
3. 동질성 검정
 - 각 그룹의 확률분포가 동일한지 검정

분할표 상에서 변수 간의 관계를 알아보기 위해 위와 같은 검정을 활용한다.

그러나 행과 열의 범주들 간의 관계를 파악하기 위해서는 대응분석이라는 새로운 방법이 필요하다.



Chi-square test of independence

카이제곱 검정 – 독립성 검정

$$H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$$

Expected value under H_0 : $E_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

$$\text{카이제곱 통계량: } \chi^2 = \sum \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(I-1)(J-1)}$$

	Male	Female	
Smoke	n_{11}	n_{12}	$n_{1\cdot}$
Non smoke	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$



Motivation

Correspondence Analysis(대응분석)의 목적

분할표의 행과 열의 관계를 보여주는 simple indices를 도출. 즉 두가지 범주형 변수의 연관성을 설명한다.

모든 행과 열 범주를 점으로 나타내어 그 relative position을 해석한다.

- 어떤 행 범주에 대해 어떤 열 범주가 가장 중요한지, 어떤 열 범주에 대해 어떤 행 범주가 가장 중요한지

방법: measure of association (χ^2 value)를 decompose

-> chi value로 알 수 있는 정보를 시각화

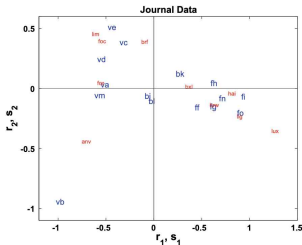
PCA와 다른 점:

PCA는 total variance를 나누는 principal components를 도출하고,

CA에서는 total chi square 값을 나누는 factor들을 구한다.

범주가 3개 이상인 경우 다중대응분석도 할 수 있다.

그러나 여기서는 범주가 2개인 경우만 살펴보자.



Motivation

Example 1

프랑스 바칼로레아 타입 & 지역의 분할표

A~H는 바칼로레아 시험 타입

(Lorraine 지역의 타입별 선호도 (conditional))

A	B	C	D	E	F	G	H
20.5	7.6	15.3	19.6	3.4	14.5	18.9	0.2

(전체 지역의 타입별 선호도 (marginal))

A	B	C	D	E	F	G	H
22.6	10.7	16.2	22.8	2.6	9.7	15.2	0.2

Lorraine 지역에서는 overall frequency에 비해

E,F,G를 선호하고, A,B,C,D를 덜 선호한다고 말할 수 있다

- 이러한 over/underrepresentation을 측정할 지표를 만들고, 각 행 범주에 대한 열 범주의 weight와 열 범주에 대한 행 범주의 weight를 부여하는 것이 CA에서 하는 일이다!



Motivation

Example 2

회사 타입과 위치의 분할표

n=3, p=3

$$\mathcal{X} = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{array}{l} \leftarrow \text{Finance} \\ \leftarrow \text{Energy} \\ \leftarrow \text{HiTech} \end{array}$$

↑ Frankfurt
↑ Berlin
↑ Munich

(conditional frequency(profile)의 weight sum)

$$s_j = c \sum_{i=1}^n r_i \frac{x_{ij}}{x_{\bullet j}}, \quad \begin{array}{l} \text{s: column weight vector} \\ \text{jth column의 average weighted frequency by r} \end{array}$$


$$r_i^* = c^* \sum_{j=1}^p s_j^* \frac{x_{ij}}{x_{i\bullet}}, \quad \begin{array}{l} \text{r: row weight vector} \\ \text{ith row의 average weighted frequency by s} \end{array}$$

r과 s를 동시에 구할 수 있다면, 이를 사용해 각 row category와 column category를 1차원 그래프에 표현할 수 있다.

그래프 상에서 r_i, s_j 가 가까운 거리에 존재하면, i행과 j열은 서로에 대해 높은 중요도를 가짐 -> positive association

그래프 상에서 r_i, s_j 가 먼 거리에 존재하면, i행과 j열은 서로에 대해 낮은 중요도를 가짐 -> negative association





5. Chi-square Decomposition

Chi-square decomposition

Measuring association by χ^2 statistic

weight vector를 계산하는 대신, 카이제곱 통계량을 decompose하여 두 변수의 연관성을 측정할 수 있다.

2차원 분할표에서 독립성 검정을 위한 카이제곱 통계량 t 는 다음과 같다.

$$t = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - E_{ij})^2 / E_{ij}, \quad E_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}.$$

x : observed value, E : expected value

$$t \sim \chi^2_{(n-1)(p-1)}$$

χ^2 decomposition은 matrix C ($n \times p$)의 SVD를 찾는 과정이다.

matrix C 의 각 element(chi value)는 독립성 가정 하에서 observed value와 expected (theoretical) value의 weighted departure라고 할 수 있다.

$$c_{ij} = (x_{ij} - E_{ij}) / E_{ij}^{1/2}.$$



Chi-square decomposition

Two ways to analyze correspondence matrix

\mathcal{X} : (unscaled) data matrix

P : Correspondence matrix ($= \frac{1}{N} \mathcal{X}$, $p_{ij} = x_{ij}/N$)

Goal: $\sum \sum \frac{p_{ij} - \hat{p}_{ij}}{ab}$ 를 minimize하는 \hat{P} 를 찾기

1. Matrix approximation method
2. Profile approximation method



Chi-square decomposition

Matrix Approximation method

\hat{P} 의 근사로 ab^T 가 많이 쓰인다.

Scaled matrix of $P: A^{-1/2} P B^{-1/2}$

$\tilde{\lambda}_k, \tilde{u}_k, \tilde{v}_k$: P 의 scaled version의 특이값과 특이벡터들

$\lambda_k = \tilde{\lambda}_{k+1}, u_k = \tilde{u}_{k+1}, v_k = \tilde{v}_{k+1}$

- rank s approximation: $\mathbf{P} = \sum_{k=1}^s \tilde{\lambda}_k (A^{1/2} \tilde{\mathbf{u}}_k) (B^{1/2} \tilde{\mathbf{v}}_k)' = \mathbf{r} \mathbf{c}' + \sum_{k=2}^s \tilde{\lambda}_k (A^{1/2} \tilde{\mathbf{u}}_k) (B^{1/2} \tilde{\mathbf{v}}_k)'$

- rank k approximation: $\mathbf{P} - \mathbf{r} \mathbf{c}' = \sum_{k=1}^K \lambda_k (\mathbf{A}^{-1/2} \mathbf{u}_k) (\mathbf{B}^{-1/2} \mathbf{v}_k)'$

=> Generalized SVD



Chi-square decomposition

Profile Approximation method

row profile, column profile $A^{-1}P, B^{-1}P$ 를 P^* 를 사용해 근사

$$(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P}^*)\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2}(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*)\mathbf{D}_c^{-1/2}$$

$$\begin{aligned}\sum_i \sum_j \frac{(p_{ij} - \hat{p}_{ij})^2}{r_i c_j} &= \sum_i r_i \sum_j \frac{(p_{ij}/r_i - p_{ij}^*)^2}{c_j} \\ &= \text{tr}[\mathbf{D}_r^{1/2}\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P}^*)\mathbf{D}_c^{-1/2}\mathbf{D}_c^{-1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P}^*)'] \\ &= \text{tr}[\mathbf{D}_r^{1/2}(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*)\mathbf{D}_c^{-1/2}\mathbf{D}_c^{-1/2}(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*)'\mathbf{D}_r^{-1/2}] \\ &= \text{tr}[(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*)\mathbf{D}_c^{-1/2}][(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*)\mathbf{D}_c^{-1/2}]' \quad (12-39)\end{aligned}$$

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \sum_{k=1}^J \tilde{\lambda}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k'$$



Chi-square decomposition

Profile Approximation method

결국 동일한 형태의 decomposition을 얻게 된다.

$$\mathbf{D}_r^{-1}\mathbf{P} = \sum_{k=1}^J \tilde{\lambda}_k \mathbf{D}_r^{-1/2} \tilde{\mathbf{u}}_k (\mathbf{D}_c^{1/2} \tilde{\mathbf{v}}_k)'$$

$$\mathbf{P}^* - \mathbf{1}_I \mathbf{c}' \doteq \sum_{k=1}^{K-1} \lambda_k \mathbf{D}_r^{-1/2} \mathbf{u}_k (\mathbf{D}_c^{1/2} \mathbf{v}_k)'$$



Chi-square decomposition

Measuring association by χ^2 statistic

2개의 범주형 변수가 $n \times p$ 의 2차원 분할표를 이룬다고 해보자.

- marginal row frequencies \mathbf{a} ($n \times 1$), marginal column frequencies \mathbf{b} ($p \times 1$) -> scaling 하는 데에 쓰임

$$\mathbf{a} = \mathbf{A} \mathbf{1}_n \quad \text{and} \quad \mathbf{b} = \mathbf{B} \mathbf{1}_p. \quad \mathbf{A} = \text{diag}(x_{i\bullet}) \quad \text{and} \quad \mathbf{B} = \text{diag}(x_{\bullet j}).$$

$$\mathbf{A} = \begin{pmatrix} x_{1\bullet} & & & \\ & x_{2\bullet} & & \\ & & \dots & \\ & & & x_{n\bullet} \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} x_{\bullet 1} & & & \\ & x_{\bullet 2} & & \\ & & \dots & \\ & & & x_{\bullet p} \end{pmatrix} \quad \mathbf{a} = \begin{pmatrix} x_{1\bullet} \\ x_{2\bullet} \\ \dots \\ x_{n\bullet} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} x_{\bullet 1} \\ x_{\bullet 2} \\ \dots \\ x_{\bullet p} \end{pmatrix}$$

$$\mathbf{C} \sqrt{\mathbf{b}} = 0 \quad \text{and} \quad \mathbf{C}^\top \sqrt{\mathbf{a}} = 0,$$



Chi-square decomposition

Measuring association by χ^2 statistic

C의 SVD: $C = \Gamma \Lambda \Delta^T$

($\Gamma: CC^T$ 의 eigenvectors, $\Delta: C^T C$ 의 eigenvectors, $\Lambda: CC^T$ 의 eigenvalues = $\text{diag}(\lambda_1^2, \dots, \lambda_R^2)$, $R = \text{rank}(C)$)

$$c_{ij} = \sum_{k=1}^R \lambda_k^{1/2} \gamma_{ik} \delta_{jk}.$$

카이제곱 통계량 t 의 decomposition은 C의 SVD, CC^T 의 고유값 분해

$$\text{tr}(CC^T) = \sum_{k=1}^R \lambda_k = \sum_{i=1}^n \sum_{j=1}^p c_{ij}^2 = t.$$

$$c_{ij} = (x_{ij} - E_{ij})/E_{ij}^{1/2}.$$

$$t = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - E_{ij})^2 / E_{ij},$$



Chi-square decomposition

Measuring association by χ^2 statistic

Duality relations에 따라, Δ, Γ 의 elements는 다음과 같다.

$$\begin{aligned}\delta_k &= \frac{1}{\sqrt{\lambda_k}} C^\top \gamma_k, \\ \gamma_k &= \frac{1}{\sqrt{\lambda_k}} C \delta_k.\end{aligned}$$

C의 행과 열의 projection

$$\begin{aligned}C \delta_k &= \sqrt{\lambda_k} \gamma_k, & \delta_k^\top \sqrt{b} &= 0, & \gamma_k^\top \sqrt{a} &= 0. \\ C^\top \gamma_k &= \sqrt{\lambda_k} \delta_k.\end{aligned}$$

이 eigenvector δ_k, γ_k 가 관심의 대상이다. χ^2 의 decomposition을 설명하는 벡터이고, 행과 열의 graphical display를 설명하는 데에 사용된다.



Chi-square decomposition

Measuring association by χ^2 statistic

($k=1\sim R$) 총 R 개의 eigenvectors, eigenvalues에서 만약 첫번째 eigenvalue가 dominant하다면 weighted departure는 아래와 같이 표현 가

$$c_{ij} = \sum_{k=1}^R \lambda_k^{1/2} \gamma_{ik} \delta_{jk} \approx \lambda_1^{1/2} \gamma_{i1} \delta_{j1}.$$

만약 γ_{i1}, δ_{j1} 가 매우 크고 같은 부호를 가지면, c_{ij} 또한 매우 크고, i 번째 행과 j 번째 열은 positive association이다.

다른 부호를 가진다면 negative association이다.

일반적으로, 첫번째 두 eigenvalues λ_1, λ_2 가 총 카이제곱 값의 대부분을 설명하고, $\gamma_1, \gamma_2, \delta_1, \delta_2$ 를 사용해 행과 열의 graphical display를 얻는다.



Chi-square decomposition

Graphical display

C의 weighted rows, weighted columns의 projection을 통해 graphical display를 표현한다.

$$\text{C의 weighted rows: } A^{-1/2}C = \begin{pmatrix} \frac{c_{11}}{\sqrt{x_{1.}}} & \dots & \frac{c_{1p}}{\sqrt{x_{1.}}} \\ \vdots & \ddots & \vdots \\ \frac{c_{n1}}{\sqrt{x_{n.}}} & \dots & \frac{c_{np}}{\sqrt{x_{n.}}} \end{pmatrix}, \text{C의 weighted columns: } B^{-1/2}C^T = \begin{pmatrix} \frac{c_{11}}{\sqrt{x_{.1}}} & \dots & \frac{c_{n1}}{\sqrt{x_{.1}}} \\ \vdots & \ddots & \vdots \\ \frac{c_{1p}}{\sqrt{x_{.p}}} & \dots & \frac{c_{np}}{\sqrt{x_{.p}}} \end{pmatrix}$$

projections on weighted rows and columns:

$$\begin{aligned} r_k &= A^{-1/2}C\delta_k = \sqrt{\lambda_k}A^{-1/2}\gamma_k, \\ s_k &= B^{-1/2}C^T\gamma_k = \sqrt{\lambda_k}B^{-1/2}\delta_k. \end{aligned}$$

marginal frequency로 정의된 natural weights(a,b)에 의해 projection은 centered at zero

$$\begin{aligned} r_k^T a &= 0, \\ s_k^T b &= 0. \end{aligned}$$



Chi-square decomposition

Graphical display

Duality relation에 의한 δ_k, γ_k 에 따르면 r, s 는 다음과 같이 표현된다.

By

$$\begin{aligned}\delta_k &= \frac{1}{\sqrt{\lambda_k}} \mathcal{C}^\top \gamma_k, \\ \gamma_k &= \frac{1}{\sqrt{\lambda_k}} \mathcal{C} \delta_k.\end{aligned}$$

Duality relation에 의한 δ_k, γ_k 에 따르면 r, s 는 다음과 같이 표현된다.

By

$$\begin{aligned}r_k &= \frac{1}{\sqrt{\lambda_k}} \mathcal{A}^{-1/2} \mathcal{C} \mathcal{B}^{1/2} s_k, \\ s_k &= \frac{1}{\sqrt{\lambda_k}} \mathcal{B}^{-1/2} \mathcal{C}^\top \mathcal{A}^{1/2} r_k,\end{aligned} \quad \Rightarrow \quad \begin{aligned}r_k &= \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}} \mathcal{A}^{-1} \mathcal{X} s_k, \\ s_k &= \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}} \mathcal{B}^{-1} \mathcal{X}^\top r_k.\end{aligned}$$

Chi-square decomposition으로 구한 projection이 앞장에서 정의한 weight vector와 동일한 관계를 가지게 됨을 알 수 있다.

$$s_j = c \sum_{i=1}^n r_i \frac{x_{ij}}{x_{\bullet j}}, \quad r_i^* = c^* \sum_{j=1}^p s_j^* \frac{x_{ij}}{x_{i\bullet}},$$



Chi-square decomposition

Graphical display

Row factors, Column factors

$$r_k = \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}} \mathcal{A}^{-1} \mathcal{X} s_k,$$
$$s_k = \sqrt{\frac{x_{\bullet\bullet}}{\lambda_k}} \mathcal{B}^{-1} \mathcal{X}^\top r_k.$$

Mean and Variance of factors

$$\bar{r}_k = \frac{1}{x_{\bullet\bullet}} r_k^\top a = 0,$$
$$\bar{s}_k = \frac{1}{x_{\bullet\bullet}} s_k^\top b = 0,$$

$$\text{Var}(r_k) = \frac{1}{x_{\bullet\bullet}} \sum_{i=1}^n x_{i\bullet} r_{ki}^2 = \frac{r_k^\top \mathcal{A} r_k}{x_{\bullet\bullet}} = \frac{\lambda_k}{x_{\bullet\bullet}},$$
$$\text{Var}(s_k) = \frac{1}{x_{\bullet\bullet}} \sum_{j=1}^p x_{\bullet j} s_{kj}^2 = \frac{s_k^\top \mathcal{B} s_k}{x_{\bullet\bullet}} = \frac{\lambda_k}{x_{\bullet\bullet}}.$$

$\frac{\lambda_k}{\sum \lambda_i}$ (t의 decomposition의 k번째 factor) 는 k번째 factor에 의한 분산의 일부로도 해석 가능



Chi-square decomposition

Absolute contributions to the variance of the factor

$$\begin{aligned}\text{Var}(r_k) &= \frac{1}{x_{\bullet\bullet}} \sum_{i=1}^n x_{i\bullet} r_{ki}^2 = \frac{r_k^\top A r_k}{x_{\bullet\bullet}} = \frac{\lambda_k}{x_{\bullet\bullet}}, \\ \text{Var}(s_k) &= \frac{1}{x_{\bullet\bullet}} \sum_{j=1}^p x_{\bullet j} s_{kj}^2 = \frac{s_k^\top B s_k}{x_{\bullet\bullet}} = \frac{\lambda_k}{x_{\bullet\bullet}}.\end{aligned}$$

absolute contributions of row i to the variance of the factor r_k :

$$C_a(i, r_k) = \frac{x_{i\bullet} r_{ki}^2}{\lambda_k}, \text{ for } i = 1, \dots, n, k = 1, \dots, R$$

어떤 행 범주가 k th row factor의 dispersion에서 가장 중요한지 알 수 있다.

absolute contributions of column j to the variance of the factor r_k :

$$C_a(j, s_k) = \frac{x_{\bullet j} s_{kj}^2}{\lambda_k}, \text{ for } j = 1, \dots, p, k = 1, \dots, R$$





6. Interpreting with Biplots

Notions

용어 정리

graphical representation: r_k, s_k

profile: 행 또는 열의 conditional frequency distribution. (profile을 projection해서 r, s 도출)

두 행 혹은 두 열의 proximity: similar profile을 가지는가?

한 행과 한 열의 proximity: 이 행(또는 열)이 특별히 important weight를 그 열(또는 행)에 가지는가?

origin: r_k, s_k 의 average. 행, 열 범주를 projection시킨 point가 origin에 가깝게 위치하면 average profile

absolute contribution: factor들의 분산 안에서 각 행 또는 열의 weight를 평가



Biplots

biplot이란

행과 열을 low dimension에 점들로 represent한 그림

지금까지 행렬을 분해하고 projection을 한 것은 결국 biplot으로 display하기 위한 것

lower dimensional factorial variables의 스칼라곱으로 해석되고, data matrix의 각 elements를 이 스칼라곱들을 통해 approximately recover하고자 함.

예를 들어, 10×5 data matrix가 있다고 하자. biplot은 10개의 row points와 5개의 column points를 찾아 50개의 스칼라곱을 만들 수 있다.

50개의 data elements에 근사할 수 있는 것을 만드는 것이 목표. row points, column points는 $q_i \in R^k, t_j \in R^k$. 보통 $k = 2$.

ex. q_7, t_4 의 스칼라곱 $\rightarrow x_{74}$ 에 근사

$$\begin{aligned}x_{ij} &= q_i^\top t_j + e_{ij} \\ &= \sum_k q_{ik} t_{jk} + e_{ij}.\end{aligned}$$



Biplots

Link between correspondence analysis and biplot

row, column frequency에 대해 x_{ij} 를 표현하면

$$x_{ij} = E_{ij} \left(1 + \frac{\sum_{k=1}^R \lambda_k^{\frac{1}{2}} \gamma_{ik} \delta_{jk}}{\sqrt{\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}}} \right)$$

$$c_{ij} = (x_{ij} - E_{ij}) / E_{ij}^{1/2}.$$

$$c_{ij} = \sum_{k=1}^R \lambda_k^{1/2} \gamma_{ik} \delta_{jk}.$$

$$E_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}.$$



Biplots

Link between correspondence analysis and biplot

profile : conditional frequencies

row profile – average row profile:

$$\left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) = \sum_{k=1}^R \lambda_k^{\frac{1}{2}} \gamma_{ik} \left(\sqrt{\frac{x_{\bullet j}}{x_{i\bullet} x_{\bullet\bullet}}} \right) \delta_{jk} = \sum_{k=1}^K \left(\frac{x_{i\bullet}}{\sqrt{\lambda_k x_{\bullet\bullet}}} r_{ki} \right) s_{kj} + e_{ij}$$

projection term 개수를 K개로 제한(보통 2)
eigenvector와 projection의 관계 사용해 정리

column profile – average column profile:

$$\left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}} \right) = \sum_{k=1}^R \lambda_k^{\frac{1}{2}} \gamma_{ik} \left(\sqrt{\frac{x_{i\bullet}}{x_{\bullet j} x_{\bullet\bullet}}} \right) \delta_{jk} = \sum_{k=1}^K \left(\frac{x_{\bullet j}}{\sqrt{\lambda_k x_{\bullet\bullet}}} s_{kj} \right) r_{ki} + e'_{ij}$$

=> column factor s_k 와 row factor r_k 의 rescaled version이 row profile과 average의 difference가 biplot을 구성한다.

row factor r_k 와 column factor s_k 의 rescaled version이 column profile과 average의 difference가 biplot을 구성한다.



Example

Belgium regions and newspapers

벨기에는 프랑스어와 네덜란드어를 공용어로 사용하고, 지역에 따라 사용하는 언어가 다르다.

row: 15개 (신문 종류) (사용 언어에 따라 3종류로 대분류할 수 있음)

column: 10개 (지역) (Flanders, Wallonia, Brussels 3 지역으로 대분류할 수 있음)

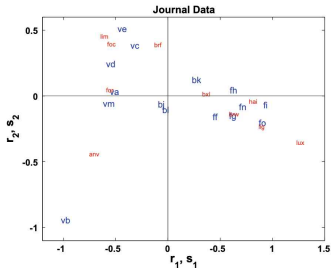


Table 15.1 Eigenvalues and percentages of the variance, Example 15.3

λ_j	Percentage of variance	Cumulated percentage
183.40	0.653	0.653
43.75	0.156	0.809
25.21	0.090	0.898
11.74	0.042	0.940



Example

Belgium regions and newspapers

Table 15.2 Absolute contributions of row factors r_k

	$C_a(i, r_1)$	$C_a(i, r_2)$	$C_a(i, r_3)$
v_a	0.0563	0.0008	0.0036
v_b	0.1555	0.5567	0.0067
v_c	0.0244	0.1179	0.0266
v_d	0.1352	0.0952	0.0164
v_e	0.0253	0.1193	0.0013
f_f	0.0314	0.0183	0.0597
f_g	0.0585	0.0162	0.0122
f_h	0.1086	0.0024	0.0656
f_i	0.1001	0.0024	0.6376
b_j	0.0029	0.0055	0.0187
b_k	0.0236	0.0278	0.0237
b_l	0.0006	0.0090	0.0064
v_m	0.1000	0.0038	0.0047
f_n	0.0966	0.0059	0.0269
f_o	0.0810	0.0188	0.0899
Total	1.0000	1.0000	1.0000

Table 15.3 Absolute contributions of column factors s_k

	$C_a(j, s_1)$	$C_a(j, s_2)$	$C_a(j, s_3)$
brw	0.0887	0.0210	0.2860
bxl	0.1259	0.0010	0.0960
anv	0.2999	0.4349	0.0029
brf	0.0064	0.2370	0.0090
foc	0.0729	0.1409	0.0033
for	0.0998	0.0023	0.0079
hai	0.1046	0.0012	0.3141
lig	0.1168	0.0355	0.1025
lim	0.0562	0.1162	0.0027
lux	0.0288	0.0101	0.1761
Total	1.0000	1.0000	1.0000



END