

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.

$$H_0: P_i = P_{i0}$$

2. 독립성 검정

- 두 변수가 서로 독립인지 검정

$$H_0: P(X_1, Y) = P(X)P(Y)$$

3. 동질성 검정

- 각 그룹의 확률분포가 동일한지 검정

$$H_0: P_{1j} = P_{2j}$$

	Y_1	Y_2
X_1	O_{11}	O_{12}
X_2	O_{21}	O_{22}

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

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

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

\uparrow Frankfurt
 \uparrow Berlin
 \uparrow Munich

(conditional frequency(profile)의 weight sum)

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

s: column weight vector

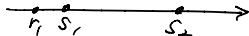
jth column의 average weighted frequency by r

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

r: row weight vector

ith row의 average weighted frequency by s

$$S_1 = c \left(\frac{r_1 \cdot 4 + r_2 \cdot 0 + r_3 \cdot 1}{x_{\bullet 1} = 5} \right) = c \frac{4r_1 + r_3}{5} \quad \begin{array}{l} S_1, S_2, S_3 \\ r_1, r_2, r_3. \end{array}$$




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 $P: A^{-1/2} P B^{-1/2} \Rightarrow U \Lambda V^T$

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

$$\tilde{\lambda}_k = \tilde{\lambda}_{k+1}, \tilde{u}_k = \tilde{u}_{k+1}, \tilde{v}_k = \tilde{v}_{k+1}$$

- rank s approximation: $P = \sum_{k=1}^s \tilde{\lambda}_k (A^{1/2} \tilde{u}_k) (B^{1/2} \tilde{v}_k)' = ab' + \sum_{k=2}^s \tilde{\lambda}_k (A^{1/2} \tilde{u}_k) (B^{1/2} \tilde{v}_k)'$ \tilde{u}_1, \tilde{v}_1 은 $\tilde{\lambda}_1$ 에 해당하는 특이벡터

- rank k approximation: $P - ab' = \sum_{k=1}^K \lambda_k (A^{-1/2} u_k) (B^{-1/2} v_k)'$ $A^{-1/2} u_1 v_1' B^{-1/2} = A^{-1/2} (A^{1/2} 1_n) (1_p B^{1/2}) B^{-1/2} = ab'$

=> Generalized SVD

$\rightarrow u_k, v_k$ 는 $A^{-1/2} (P - ab') B^{-1/2}$ 의 eigenvector

$$(A^{-1/2} u_k)' A^{-1} (A^{-1/2} u_k) = u_k' u_k = 1$$

$$(B^{1/2} v_k)' B^{-1} (B^{1/2} v_k) = v_k' v_k = 1 \Rightarrow \text{SVD}$$

(rank s approximation)

$$D = A^{-1/2} P B^{-1/2} = \sum_{k=1}^s \tilde{\lambda}_k \tilde{u}_k \tilde{v}_k'$$

$$\hat{P} = A^{1/2} D B^{1/2} = \sum_{k=1}^s \tilde{\lambda}_k (A^{1/2} \tilde{u}_k) (B^{1/2} \tilde{v}_k)'$$

with error $P - \hat{P} = \sum_{k=s+1}^p \tilde{\lambda}_k$

ab' : rank 1 approximation

$$\rightarrow \tilde{u}_1 = A^{1/2} 1_n, \tilde{v}_1 = B^{1/2} 1_p \text{ 이고 이 때,}$$

$$\tilde{u}_1' (A^{-1/2} P B^{-1/2}) = (A^{1/2} 1_n)' (A^{-1/2} P B^{-1/2}) = 1_n' P B^{-1/2} = \tilde{v}_1', (A^{-1/2} P B^{-1/2}) \tilde{v}_1 = \tilde{u}_1$$



Chi-square decomposition

$$D_r = A \quad D_c = B$$

Profile Approximation method

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

$$(D_r^{-1}P - P^*)D_c^{-1/2} = D_r^{-1/2}(D_r^{-1/2}P - D_r^{1/2}P^*)D_c^{-1/2}$$

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

$$D_r^{-1/2}PD_c^{-1/2} = \sum_{k=1}^J \tilde{\lambda}_k \tilde{u}_k \tilde{v}_k'$$



Chi-square decomposition

$$D_r^{-1/2} \times (D_r^{-1/2} P D_c^{-1/2}) \times D_c^{1/2} = \sum_{k=1}^J \lambda_k (D_r^{-1/2} \tilde{u}_k) (D_c^{1/2} \tilde{v}_k)'$$

Profile Approximation method

결국 동일한 형태의 decomposition을 얻게 된다.

$$D_r^{-1} P = \sum_{k=1}^J \tilde{\lambda}_k \overset{A^{1/2}}{D_r^{-1/2}} \tilde{u}_k (\overset{B^{1/2}}{D_c^{1/2}} \tilde{v}_k)'$$

$$\tilde{u}_1 = A^{1/2} \mathbf{1}_n, \quad \tilde{v}_1 = B^{1/2} \mathbf{1}_p$$

$$A^{-1/2} \tilde{u}_1 = A^{-1/2} (A^{1/2} \mathbf{1}_n) = \mathbf{1}_n$$

$$B^{1/2} B^{1/2} \mathbf{1}_p = b.$$

$$P^* - \mathbf{1}_p b \doteq \sum_{k=1}^{K-1} \lambda_k D_r^{-1/2} \tilde{u}_k (D_c^{1/2} \tilde{v}_k)'$$

$\mathbf{1}_p b$: leading term of decomposition



Chi-square decomposition

Measuring association by χ^2 statistic

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

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

$$a = A1_n \text{ and } b = B1_p. \quad A = \text{diag}(x_{i\bullet}) \text{ and } B = \text{diag}(x_{\bullet j}).$$

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

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

$$C\sqrt{b} = 0 \text{ and } C^T\sqrt{a} = 0,$$

($C\sqrt{b} = 0$ proof.)

$$C_{ij}\sqrt{b_j} = \left(\frac{x_{ij} - E_{ij}}{\sqrt{E_{ij}}} \right) \sqrt{x_{\bullet j}}$$

$$\left(E_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{N} \right) = \frac{N x_{ij} - x_{i\bullet} x_{\bullet j}}{\sqrt{x_{i\bullet}}} \frac{1}{\sqrt{N}}$$

Sum over all j

$$C\sqrt{b} = \frac{N x_{i\bullet} - x_{i\bullet} N}{\sqrt{x_{i\bullet}} \sqrt{N}} = 0$$



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.

$$\begin{aligned} \Delta \rightarrow \delta_k &= \frac{1}{\sqrt{\lambda_k}} C^T \gamma_k, \\ \Gamma \rightarrow \gamma_k &= \frac{1}{\sqrt{\lambda_k}} C \delta_k. \end{aligned} \quad C \delta_k = \frac{1}{\sqrt{\lambda_k}} \underbrace{(C C^T)}_{\lambda_k} \gamma_k = \sqrt{\lambda_k} \gamma_k$$

C의 행과 열의 projection ↗

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

($\delta_k^T \sqrt{b} = 0$ proof)

$$C \sqrt{b} = \Gamma \Lambda \Delta^T \sqrt{b} = 0 \quad (\sqrt{b} = 0 \text{ is known})$$

$$\Gamma^T (\Gamma \Lambda \Delta^T) \sqrt{b} = \Lambda \Delta^T \sqrt{b} = 0$$

$$\sqrt{\lambda_k} (\delta_k^T \sqrt{b}) = 0$$

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:

$$r_k = A^{-1/2}C\delta_k = \sqrt{\lambda_k}A^{-1/2}\gamma_k,$$

$$s_k = B^{-1/2}C^T\gamma_k = \sqrt{\lambda_k}B^{-1/2}\delta_k.$$

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

$$\frac{r_k^T a = 0,}{s_k^T b = 0.}$$

$$(h_k^T a = 0 \text{ proof})$$

$$A^{-1/2} \cdot a = A^{-1/2} \cdot A \cdot 1_n = A^{1/2} 1_n = \sqrt{a}$$

$$(\sqrt{\lambda_k} (A^{-1/2} \gamma_k)^T) a$$

$$= \sqrt{\lambda_k} \cdot \gamma_k^T A^{-1/2} \cdot a$$

$$= \sqrt{\lambda_k} \cdot \gamma_k^T \cdot \sqrt{a} = 0$$

$$(\because \gamma_k^T \sqrt{a} = 0)$$



Chi-square decomposition

Graphical display

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

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

파인한,
v

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

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 x 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

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

$$c_{ij} = (x_{ij} - E_{ij}) / E_{ij}^{1/2} \rightarrow \sqrt{E_{i\bullet}} c_{i\bullet} + E_{\bullet j} = x_{i\bullet}$$

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

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



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)

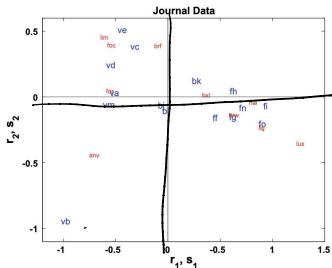


Table 15.1 Eigenvalues and percentages of the variance, Example 15.3

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

centered 0



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