MDS 개요

- Dimension Reduction
  - (n×p) 데이터 행렬의 차원 축소를 통해 내재된 구조 발견
  - 변수 p 축소
    - 주성분 분석 (Y=LX): 원변수 선형결합, 선형계수 L (loading 값)
    - 요인분석 (X=LF): 원변수 그룹, 요인점수(F)
  - 개체 n 축약
    - 군집분석: 유사성(거리) 개념으로 개체 군집화

- 범주형 데이터 차원 축약?
  - 변수 축약
    - 변수들간의 연관관계는 교차표(분할표)에 의해서 표현되므로 상관관계 값에 의한 변수의 축약은 가능하지 않음
  - 개체 축약
    - 개체간의 유사성(거리)을 측정할 수 있나? 범주형 변수의 범주(가질 수 있는 값)는 빈도로만 정리할 수 있는데...

- MDS 개념
  - n개의 개체를 2차원 가시적 공간에 나타내는 방법
  - 각 개체간 유사성(similarity) 혹은 거리는 저차원으로 옮겨지더라도 원래 유사성 크기를 갖는다.
  - 유사성 개념
    - 개체를 저 차원 가시적 공간(2차원)에 나타내려면 각 개체간 거리(유사성)를 측정해야 한다.
    - MDS는 개체(행), 변수(열) 모두 저차원 공간 표현 가능

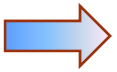
- 개체간 유사성 측정
  - metric 방법
    - Euclidean distance ► (측정형 변수 거리)
    - 각 개체의 유사성(거리)을 사람들이 리커드 척도나 순위 평가
    - 유사성을 계산하여 개체를 분류하는 면에서는 군집분석과 유사
    - 군집분석은 개체를 군집화 하고, 주성분 점수에 의해 개체를 표현하나, MDS는 유사성에 의해 단지 2차원 공간에 표현
  - non-Metric 방법
    - 평가자 들이 개체를 주관적으로 분류하게 하고 그로부터 얻어지는 <u>빈도</u>로부터 유사성을 측정

MDS (개념, 페이지 278)

- 개체를 저차원 공간에 표현
  - 개체의 유사성(similarity)의 상대적 크기를 2-3차원 공간에 표현

- 유사성 측정
  - Metric 방법
    - 개체 (i, j) 유사성: 거리 개념
    - 변수 유형 $S_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$
      - 개체 속성을 측정하는 변수: 군집분석의 변량과 동일
      - 리커트 척도, 우선 순위: (회사 1, 회사2, 회사3, ...) 혹은 (속성1, 속성2, 속성3, ...)
      - ($p \leq 2$)일 때도 유사성(거리)을 계산
  - Non-metric 방법
    - 개체에 대한 빈도표를 이용
    - (상대)빈도(f_{ij})가 개체간 유사성 (S_{ij}) 측정

개체 \ 변수	X_1	X_2	⋯→	X_p
1	x_{11}	x_{12}	⋯→	x_{1p}
2	x_{21}	x_{22}	⋯→	x_{2p}
⋮	⋮	⋮	⋯→	⋮
n	x_{n1}	x_{n2}	⋯→	x_{np}



- 예제 (non-metric)
  - 유사 상품 분류: 마케팅 전문가 15명에게 n개 상품을 주관적인 판단에 의해 임의로 분류하게 한 후 상대빈도표(?/15) 작성
    - 예: (상품 (1, 2) 유사성 계산) 상품 (1, 2)를 동일 군집으로 분류한 전문가가 10명이라면 유사성은 10/15가 된다.

- 예제 (metric)
  - Under-arm Deodorant생산 회사에서 판매 전략을 세우기 위하여 각 제품들이 서로 얼마나 유사한지(가까운지) 알아보려고 한다. 이를 위하여 소비자를 임의로 선택하여 제품의 각 분야(향기, 냄새 제거 정도, 사용 편리 정도, 옷에 묻어나는 정도)를 10점 만점으로 평가
    - 고객 평가점수에 의한 제품의 유사성 정도를 2차원 공간에 표현
    - 제품 평가 유사성에 따른 고객 세분화

개체 \ 개체	1	2	⋯→	n
1	0			
2	S_{21}	0		
⋮	⋮	⋮	0	
n	S_{n1}	S_{n2}	⋯→	0

MDS (알고리즘, 페이지 280)

- 표현 방법
  - 각 개체간 유사성(거리)을 측정한다.
  - 개체의 개수가 n개인 경우 k=n(n-1)/n개 유사성 그룹이 존재한다.
  - 유사성이 작은 것부터 크기 순으로 배열한다.
    - 이를 이용하여 개체를 m(= 2)차원으로 공간으로 줄일 경우 개체간의 거리를 구한다.
  - 임의의 한 좌표에 한 개체를 표현하고 나머지 개체들은 상대적 유사성을 고려하여 좌표에 표현한다.
- Stress 값
  - 2차원 공간으로 줄일 수 있는지를 알아 보는 측정치
  - S^r은 차원이 2차원으로 줄었을 때 개체의 유사성

$$stress = \frac{[\sum_{i<j}(S_{ij} - S^r_{ij})^2 / S_{ij}]}{\sum_{i<j}(S_{ij})}$$

Stress	Goodness of fits
20%	Poor
10%	Fair
5%	Good
2.5%	Excellent

- 차원(dimension)과 위치(coordination)의 의미
  - 아무 의미 없음
  - 개체 간 유사성은 얼마나 가까이 있나에 의해 해석됨

- 예제 데이터 CITY.xls
  - 미국 도시간 거리(도시간 유사성)
  - 개체 유사성 행렬 (similarity matrix)

ID	Atlanta	Chicago	Denver	Houston	LA	Miami	NY	SF	Seattles	DC
Atlanta	0									
Chicago	567	0								
Denver	1212	920	0							
Houston	701	940	879	0						
LA	1936	1745	831	1374	0					
Miami	604	1188	1726	968	2339	0				
NY	748	713	1631	1420	2451	1092	0			
SF	2139	1858	949	1645	347	259	2571	0		
Seattles	2128	1737	1021	1891	959	2734	2408	678	0	
DC	543	597	1494	1220	2300	923	205	2442	2329	0

MDS 예제 (페이지 285)

▪R 프로그램

```
> city=read.table("도시거리.csv",header=T,
+ sep=",",na.string=".")
> citys=city[2:10,2:10]
> citys
   Altanta Chicago Denver Houston   LA Miami   NY
2      567      NA     NA      NA   NA    NA   NA
3     1212     920     NA      NA   NA    NA   NA
4      701     940    879      NA   NA    NA   NA
5     1936    1745    831    1374   NA    NA   NA
6      604    1188   1726     968 2339    NA   NA
7      748     713   1631    1420 2451  1092   NA
8     2139    1858    949    1645  347   259 2571
9     2128    1737   1021    1891  959  2734 2408
10     543     597   1494    1220 2300   923  205
> city.mds=cmdscale(citys,eig=TRUE, k=2)
이하에 에러cmdscale(citys, eig = TRUE, k = 2) :
  'd'에서 NA값은 허용되지 않습니다
```

- I just give up? I don't know how to delete "NA"

▪결과

```
> city=read.table("도시거리.csv",header=F,
+ sep=",",fill=T,flush=T)
> citys=city[3:11,2:10]
> citys0=citys
> fix(citys)
> city=read.table("도시거리.csv",header=F,
+ sep=",",fill=T,flush=T)
> citys=city[3:11,2:10]
> citys0=citys
> fix(citys)
```

R 자료 편집기

	row.names	V2	V3	V4	V5
1	3	567			
2	4	1212	920		
3	5	701	940	879	
4	6	1936	1745	831	1374
5	7	604	1188	1726	968
6	8	748	713	1631	1420

페이지 (페이지290)

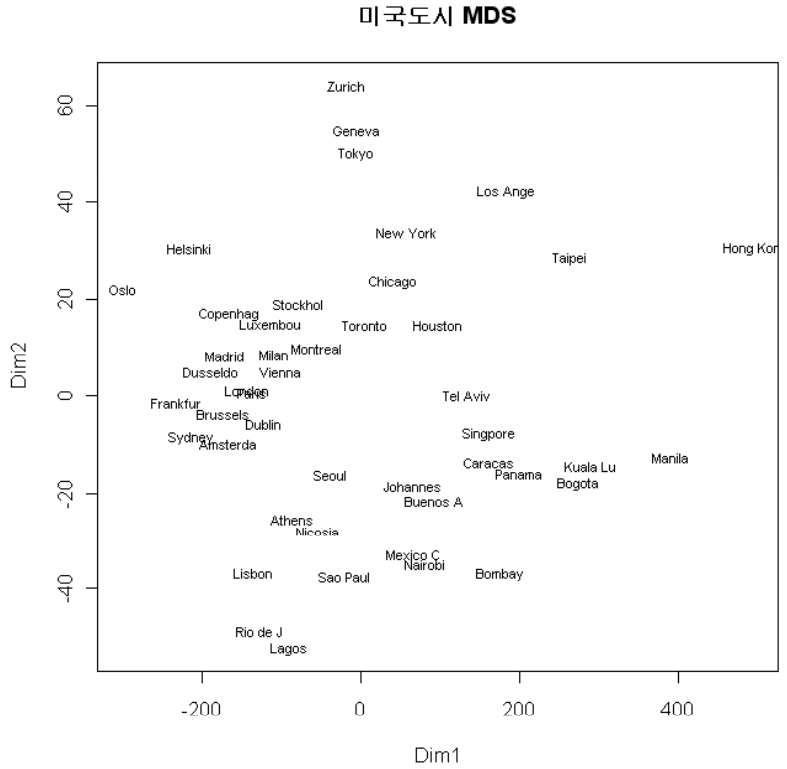
▪예제 데이터 CITY1.xls

- 경제적 변인에 의해 도시를 군집화 하려 한다. 도시 이름, 12개 직종 노동시간 가중 평균(work), 물가(price), 시간당 임금(salary)

```
> city2=read.table("도시경제.csv",header=T,
+ sep=",",na.string=".")
> city2s=city2[,2:4]
> city.d=dist(city2s)
> city.mds=cmdscale(city.d,eig=TRUE, k=2)
>
> # plot solution
> x=city.mds$points[,1]
> y=city.mds$points[,2]
> plot(x, y, xlab="Dim1", ylab="Dim2",
+ main="미국도시 MDS", type="n")
> text(x, y,labels=city2$City, cex=.7)
```

```
> city2
       City Work Price Salary
1  Amsterda 1714    66     49
2    Athens 1792    54     30
3    Bogota 2152    38     12
4    Bombay 2052    30      5
5  Brussels 1708    74     51
6  Buenos A 1971    56     13
```

▪결과




```
> city.d
          1         2         3         4
2  81.172655
3 440.450905 360.804656
4 342.747721 262.299447 100.563413
5  10.198039  88.865066 447.161045 349.839963
6 259.701752 179.816573 181.895574  85.445889
```

Correspondence Analysis 개요 (페이지 294)

▪개념

- 범주형 변수 범주의 유사성 표현
- 교차표(분할표)로 나타내어지는 자료의 행과 열 범주를 저차원 공간상(2차원)의 좌표로 표현하여 관계를 탐구하려는 탐색적 자료 분석 기법

▪기원

- 대응분석의 수리적인 기원은 1930년대 Hirshfeld의 논문『상관관계와 분할표의 연관성』
- 대응분석의 기하적인 면은 1960년대 프랑스에서 Jean-Paul Benzecri에 의해서 발전되었다.
- 일본: 1950년대 Chikio Hayashi에 의해서 수량화 제3방법으로 개발되어 발전
- 프랑스: 1960년대 Jean-Paul Benzecri가 이끄는 자료분석 모임이 다양한 분야로부터 수집된 자료를 분석하는데 대응분석 기법을 응용하고 발전

$$\text{검정통계량} = \frac{\sum_{i=1}^{k}(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

▪RXC 분할표

- π_{ij}: (X, Y) 결합밀도함수
- π_{i+}: (X) 주변밀도함수
- π_{+j}: (Y) 주변밀도함수

X \ Y	1	2	...	C	Total
1	π_{11}	π_{12}	...	π_{1c}	π_{1+}
2	π_{21}	π_{22}	...	π_{2c}	π_{2+}
...	...	...	...	...	...
R	π_{r1}	π_{r2}	...	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	...	π_{+c}	π_{++}

▪Homogeneity (동질성)

- 각 행에 대해 열의 분포가 동일한가?

▪Independence (독립성)

$$H_0 : \pi_{ij} = \pi_{kj} \text{ for } j = 1,2,\ldots,c \text{ and } k \neq i$$

- (X, Y)는 서로 독립인가?

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

▪결과 해석

- χ^2검정 결과 p-value(유의확률) 0.05보다 적으면 두 변수 상관관계 존재
- 관계 해석: 행 퍼센트 혹은 열 퍼센트에 의한 차이 해석
- R×C 셀이 많아지면 퍼센트에 의한 해석이 복잡해지고 신뢰도가 떨어짐
- 행 범주, 열 범주의 유사성 정도를 표현하지 못함

Correspondence Analysis (페이지 295)

- Notation
  - n_{ij}: (i, j) 셀 관측빈도
  - n_{i+}: (i)번째 행의 관측빈도 합
  - n_{+j}: (j)번째 열의 관측빈도 합
- 방법
  - (i, j) 셀의 빈도 n_{ij} (≥ 0)의 i번째 행 (n_{i1}, ..., n_{ic})은 총빈도가 $n_{i+} = n_{i1} + \dots + n_{ic}$ 이고 C개 범주를 갖는 다항분포
  - Multinomial 분포의 대응 확률은 상대빈도 $f_{ij} = n_{ij} / n$: 이것을 행 프로파일 (row profile)이라 정의
  - 각 행의 상대적 빈도 $f_{i+} = f_{i1} + \dots + f_{ic}$ 를 선형계수로 (주성분 분석과 유사) 하여 좌표 계산
  - $r_i = (f_{i1} / f_{i+}, f_{i2} / f_{i+}, \dots, f_{ic} / f_{i+})$는 C 차원 가중 Euclid 공간의 좌표
  - 가중(weighted) Euclid 공간이란 두 개의 좌표 r_a, r_b 사이의 거리가 다음과 같이 정의
    - $f_{+j} = f_{1j} + \dots + f_{rj}$

$$d(r_a, r_b) = \sqrt{\sum_j \left(\frac{f_{aj}}{f_{a+}} - \frac{f_{bj}}{f_{b+}}\right)^2 / f_{+j}}$$

  - 같은 방식으로 열 프로파일의 좌표 및 개체 거리 계산
  - 행, 열 프로파일을 각각 2차원 공간에 표현하거나 동시에 표현

X \ Y	1	2	...	C	Total
1	n_{11}	n_{12}	...	n_{1c}	n_{1+}
2	n_{21}	n_{22}	...	n_{2c}	n_{2+}
...	...	...	...	...	...
R	n_{r1}	n_{r2}	...	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	...	n_{+c}	n_{++}

- 예제 데이터
  - 학과별 영어등급 조사한 데이터이다.

	전공	영어등급	빈도
1	경영	A	78
2	경영	B	22
3	경영	C	20
4	경제	A	65
5	경제	B	8
6	경제	C	2
7	통계	A	68
8	통계	B	30
9	통계	C	7

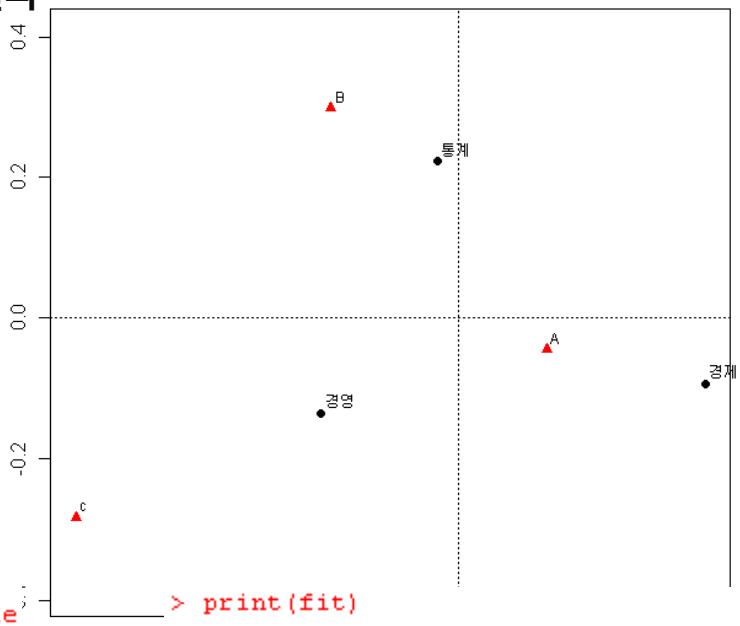
대응분석 (예제)

▪R 프로그램

```
> # Correspondence Analysis
> c1=cbind(rep(("경영"),c(78)),rep(("A"),c(78)))
> c2=cbind(rep(("경영"),c(22)),rep(("B"),c(22)))
> c3=cbind(rep(("경영"),c(20)),rep(("c"),c(20)))
> c4=cbind(rep(("경제"),c(65)),rep(("A"),c(65)))
> c5=cbind(rep(("경제"),c(8)),rep(("B"),c(8)))
> c6=cbind(rep(("경제"),c(2)),rep(("c"),c(2)))
> c7=cbind(rep(("통계"),c(68)),rep(("A"),c(68)))
> c8=cbind(rep(("통계"),c(30)),rep(("B"),c(30)))
> c9=cbind(rep(("통계"),c(7)),rep(("c"),c(7)))
> mydata=data.frame(rbind(c1,c2,c3,c4,c5,c6,c7,c8,c9))
> names(mydata)
[1] "X1" "X2"
> length(mydata$X1)
[1] 300
```

```
> library(ca)
> mytable=with(mydata, table(X1,X2)) # create a 2 way table
> prop.table(mytable, 1) # row percentages
        X2
X1                A          B          c
  경영 0.65000000 0.18333333 0.16666667
  경제 0.86666667 0.10666667 0.02666667
  통계 0.64761905 0.28571429 0.06666667
> prop.table(mytable, 2) # column percentages
        X2
X1                A          B          c
  경영 0.36966825 0.36666667 0.68965517
  경제 0.30805687 0.13333333 0.06896552
  통계 0.32227488 0.50000000 0.24137931
> fit=ca(mytable)
```

▪결과




```
> print(fit)

 Principal inertias (eigenvalues):
            1        2
Value       0.046172 0.026982
Percentage  63.12%   36.88%


 Rows:
              경영       경제       통계
Mass      0.400000   0.250000   0.350000
ChiDist   0.236902   0.363560   0.224630
Inertia   0.022449   0.033044   0.017660
Dim. 1   -0.903071   1.633900  -0.134991
Dim. 2   -0.827323  -0.574778   1.356068
```