MDS, 요인분석(factor analysis)

Jinseog Kim

October 30, 2007

1 Multidimensional scaling (MDS)

MDS는 R^p공간에 있는 자료점간의 유사도 (혹은 비유사도)를 측정하여 2차원 혹은 3차원 공간에 보여주는 통계방법론을 말한다. 즉

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p \rightarrow \mathbf{z}_1, \ldots, \mathbf{z}_n \in R^k, \ k < p.$$

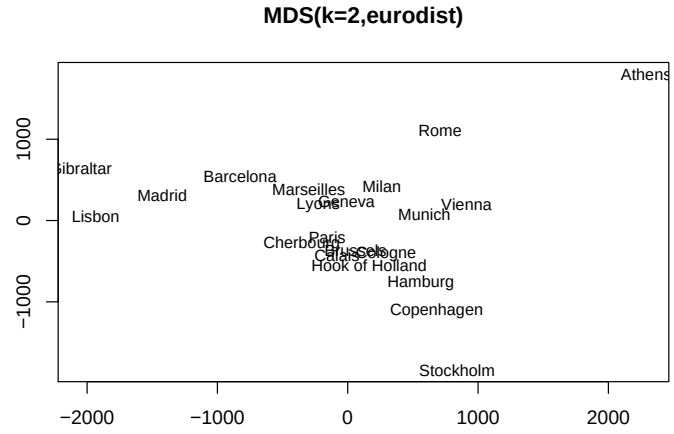
$\mathbf{x}_i$와 $\mathbf{x}_j$가 각각 R^p공간에 있는 자료점일 때, $\mathbf{x}_i$와 $\mathbf{x}_j$간의 비유사도를 $s_{\mathbf{x}}(i,j)$ 로 정의하고,

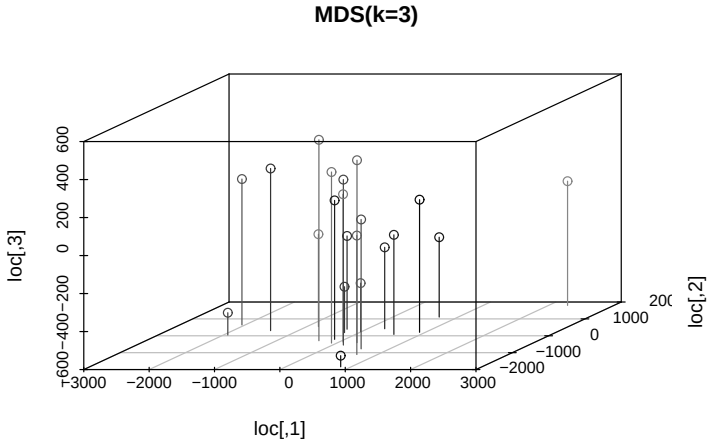
R^k에서의 $\mathbf{z}_i$와 $\mathbf{z}_j$간의 비유사도를 $s_{\mathbf{z}}(i,j)$ 정의 하면 각 자료점들의 비유사도의 차이가 가장 작게 되도록 하는 $\mathbf{z}_i, i = 1, \ldots, n$을 찾을 수 있다. 즉,

$$Stress(\mathbf{z}_1, \ldots, \mathbf{z}_n) = \sum_{i \neq j = 1, \ldots, n} \{s_{\mathbf{x}}(i,j) - s_{\mathbf{z}}(i,j)\}^2.$$

```
loc <- cmdscale(eurodist)
plot(loc, type="n", xlab="", ylab="", main="MDS(k=2,eurodist)")
text(loc, rownames(loc), cex=0.8)
```

```
loc <- cmdscale(eurodist, k=3)
library(scatterplot3d)
s3d<-scatterplot3d(loc, type="h",
    color=grey(length(loc[,1]):1/40), main="MDS(k=3)")
```





- Classical multidimensional scaling

$$s_{\mathbf{x}}(i,j) =< \mathbf{x}_i - \bar{\mathbf{x}}, \mathbf{x}_j - \bar{\mathbf{x}} >= \sum_{k=1}^{p} (x_{ik} - \bar{x}_k)(x_{jk} - \bar{x}_k)$$

$$s_{\mathbf{z}}(i,j) =< \mathbf{z}_i - \bar{\mathbf{z}}, \mathbf{z}_j - \bar{\mathbf{z}} >= \sum_{k=1}^{k} (z_{ik} - \bar{z}_k)(z_{jk} - \bar{z}_k)$$

- Metric multidimensional scaling

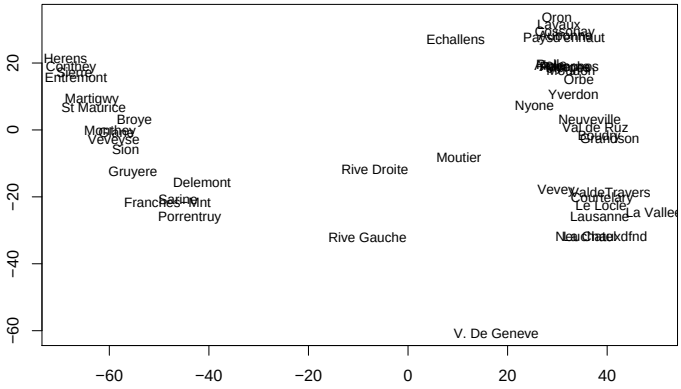
$$s_{\mathbf{x}}(i,j) = ||\mathbf{x}_i - \mathbf{x}_j|| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

- Non-metric multidimensional scaling

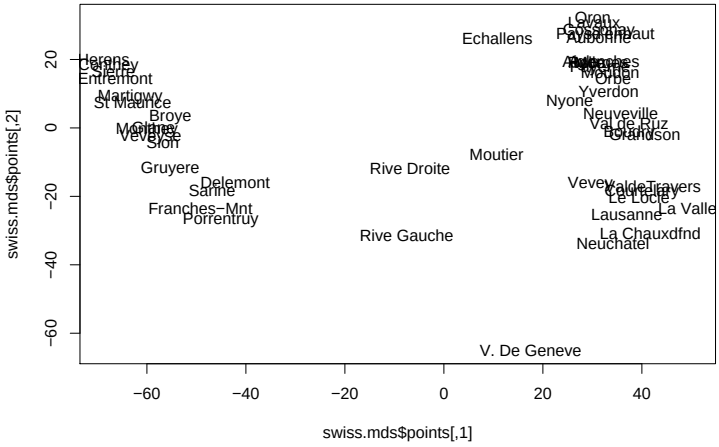
$$Stress(\mathbf{z}_1, \ldots, \mathbf{z}_n) = \sum_{i \neq j = 1, \ldots, n} \frac{\{s_{\mathbf{x}}(i,j) - s_{\mathbf{z}}(i,j)\}^2}{s_{\mathbf{x}}(i,j)}.$$

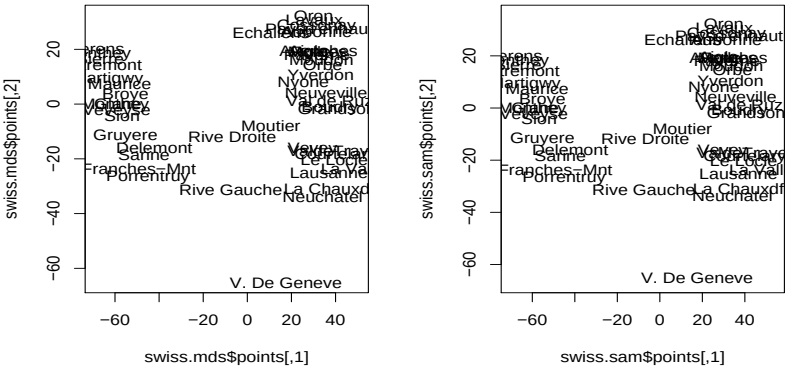
```
data(swiss)

swiss.x <- as.matrix(swiss[, -1])

loc <- cmdscale(dist(swiss.x))

plot(loc, type="n", xlab="", ylab="", main="MDS(k=2,swiss)")

text(loc, rownames(loc), cex=0.8)


par(mfrow=c(1,2))

swiss.dist <- dist(swiss.x)

swiss.mds <- isoMDS(swiss.dist)

plot(swiss.mds$points, type = "n")

text(swiss.mds$points, labels = rownames(swiss.x))


swiss.x <- as.matrix(swiss[, -1])

swiss.sam <- sammon(dist(swiss.x))

plot(swiss.sam$points, type = "n")

text(swiss.mds$points, labels = rownames(swiss.x))
```

MDS(k=2,swiss)



Non−metric





2 요인분석 (factor analysis)

p 차원 확률벡터 $\mathbf{x} = (x_1, \ldots, x_p)'$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산행렬이 Σ 일 때,

- 주성분분석 (Principal Component Analysis)은 p차원자료를 설명력이 높은 몇 개의 차원($q < p$)으로 축소하기 위한 분석방법이다.

$$\mathbf{y} = P'\mathbf{x}.$$

- 요인분석은 p차원자료를 적은 수($k < p$)의 잠재변수(latent factor)로 설명하는 분석방법이다. k-요인모형은 다음과 같다.

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \boldsymbol{\epsilon} \ ,$$

이 때, $\Lambda = \{\lambda_{ij}\}$는 $p \times k$ 행렬이고 λ_{ij}를 인자적재값(factor loading)이라고 부른다. 여기서 $\mathbf{f}$는 관측할 수 없는 변수이고, 이를 잠재요인(latent factor) 혹은 공통요인(common factor)라고 한다. 또한 오차항 $\boldsymbol{\epsilon}$의 원소를 유일인자 혹은 특정인자(unique factor or specific factor)라고 한다.

요인분석모형의 가정은 아래와 같다.

A1 $E(\mathbf{f}) = \mathbf{0}$, $Cov(\mathbf{f}) = \mathbf{I_k}$

A2 $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Cov(\boldsymbol{\epsilon}) = \Psi = \text{diag}(\psi_1, \ldots, \psi_p)$

A3 $\mathbf{f}$, $\boldsymbol{\epsilon}$는 비상관(uncorrelated)되어 있다.

이 때, $\mathbf{x}$의 공분산은

$$Cov(\mathbf{x}) = \Sigma = \Lambda\Lambda' + \Psi$$

로 표현할 수 있다.

위의 식에서 각 변수(x_i)의 분산은 양변의 대각원소를 비교하여,

$$\begin{aligned} \sigma_{ii} &= \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i = h_i^2 + \psi_i \\ &= \text{ comumunality + specific variance(or uniqueness)} \end{aligned}$$

와 같이 나타내어 진다.

이 때 첫번째 성분 $\sum_{j=1}^{k} \lambda_{ij}^2$를 공통성(communality)라고 부르며, 이는 x_i와 잠재인자들이 공유하는 분산을 의미한다. 두번째 성분인 ψ_i는 유일분산(unique variance), 혹은 특정분산(specific variance)라고 부르며 변수 잠재인자가 설명하지 못하는 x_i의 분산, 즉 오차이다.

2.1 인자적재값의 비유일성

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \boldsymbol{\epsilon} \ ,$$

모형에서의 $\mathbf{x}$의 공분산은

$$Cov(\mathbf{x}) = \Sigma = \Lambda\Lambda' + \Psi$$

이다.

만일 Γ를 다음을 만족하는 직교행렬(orthogonal matrix)라고 하자.

$$\Lambda_* = \Lambda\Gamma, \ \mathbf{f}_* = \Gamma'\mathbf{f}$$

그러면,

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda_*\mathbf{f}_* + \boldsymbol{\epsilon} \ ,$$

모형하에서의 공분산 행렬은

$$\Lambda_*\Lambda_*' + \Psi = \Lambda\Gamma\Gamma'\Lambda' + \Psi = \Lambda\Lambda' + \Psi.$$

따라서 Λ를 유일하게 결정하기 위한 방법은?(이를 identifiability condition 이라고 한다)

$$\Lambda' W^{-1}\Lambda = \text{대각행렬}.$$

여기서 W는 다음 네가지 경우를 생각할 수 있다.

$$W = \begin{cases} \Psi \\ I_k \\ \Sigma \\ \text{diag}(\Sigma) \end{cases}$$

2.2 인자분석의 절차

1. 어떤 k에 대하여 모수를 추정한다.

2. 적합도 검정을 시행한 후 인자를 회전하여 해석이 용이한 인자적재값을 찾는다.

3. 잠재인자값(Factor score)을 추정한다.

2.3 모수의 추정

인자분석에서 추정모수는

- $\boldsymbol{\mu}$: 평균벡터
- Λ : Factor loading
- Ψ : 오차항의 분산

모수추정은 표본에서 얻어진 평균벡터($\boldsymbol{\mu}$) 및 표본공분산행렬(S_n)을 이용하며 기본적인 아이디어는 다음과 같다.

$$\boldsymbol{\mu} = \bar{\mathbf{x}}$$

$$S_n = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$$

이때, S_n의 독립된 원소의 개수는 $p(p+1)/2$이고, Λ와 Ψ의 모수의 개수는 $pk+p = p(k+1)$이다. 또한 제약조건의 수는 $k(k-1)/2$이다. 따라서 미지인 모수와 방정식의 수간의 관계를 이용하면

$$s = \frac{1}{2}\{(p-k)^2 - (p-k)\}$$

이고, $s \geq 0$인 경우 방정식의 해를 구할 수 있다. 다만 $s = 0$인 경우, factor model이 더 간단한 것이 아니므로 우리는 $s > 0$인 경우에만 factor analysis를 하는 것이 의미가 있다.

- 최우추정법(maximum likelihood method): 다변량 정규분포의 가정과 유일해 조건으로 부터 얻어짐.

- 주성분분석법(principal factor method):

$$\hat{\Lambda}\hat{\Lambda}' = S_n - \Psi = GLG' \text{ by spectral decomposition.}$$

$$\hat{\Lambda}\hat{\Lambda}' = GL^{1/2}L^{1/2}G' = GL^{1/2}(GL^{1/2})'$$

따라서 $\hat{\Lambda} \stackrel{?}{=} GL^{1/2}$로 생각할 수 있으나, $\hat{\Lambda}$은 $p \times k$ matrix인 반면 $GL^{1/2}$ 은 $p \times p$ 행렬이므로 옳지않다. 여기서 우리는 G_1를 k번째까지 큰 eigen values $l_1 \geq l_2 \geq \ldots \geq l_k$들을 모아 재정렬할 때 그와 대응되는 eigen vector들로 만들어진 행렬로 하고 L_1 역시 같은 방법으로 eigen-value들을 대각원소로 하는 행렬로 구성하자. 그러면

$$\hat{\Lambda} = G_1 L_1^{1/2}$$

이 된다.

2.4 적합도 검정 및 잠재인자수의 결정

잠재인자수를 결정하는 방법으로는 다음과 같은 방법을 이용할 수 있다.

- 전체분산에 대한 k개의 factors로 얻어지는 분산의 비로 추정, 이를테면 80%이상.
- eigen value가 평균 eigen value보다 큰 값에서 선택.
- Using scree plot
- (Mardia, Kent, and Bibby (1979, pg. 258), Goodness of Fit)다음의 가설을 우도비 검정통계량을 이용하여 검정

$$H_k : \Sigma = \Lambda\Lambda' + \Psi,$$

$$n - \frac{2p + 4k + 11}{6} \log\left(\frac{|\hat{\Lambda}\hat{\Lambda}'\hat{\Psi}|}{|S|}\right) \approx \chi^2(\nu)$$

where $\nu = \frac{1}{2}[(p-k)^2 - p - k]$

2.5 잠재인자값의 추정

- 회귀분석방법:

$$\mathbf{x} - \boldsymbol{\mu} \sim N_p(\Lambda\mathbf{f}, \boldsymbol{\Psi})$$

임을 이용하면, 최소제곱추정량은

$$\mathbf{f} = (\Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

이다.

- Bartlette 방법

3 R-example

```
v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)

v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)

v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)

v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)

v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)

v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)

m1 <- cbind(v1,v2,v3,v4,v5,v6)

cor(m1)

          v1        v2        v3        v4        v5        v6

v1 1.0000000 0.9393083 0.5128866 0.4320310 0.4664948 0.4086076

v2 0.9393083 1.0000000 0.4124441 0.4084281 0.4363925 0.4326113

v3 0.5128866 0.4124441 1.0000000 0.8770750 0.5128866 0.4320310

v4 0.4320310 0.4084281 0.8770750 1.0000000 0.4320310 0.4323259

v5 0.4664948 0.4363925 0.5128866 0.4320310 1.0000000 0.9473451

v6 0.4086076 0.4326113 0.4320310 0.4323259 0.9473451 1.0000000
```

3.1 No rotation

```
> factanal(m1, factors=3, rotation="none") # no rotation
```

```
Call:
factanal(x = m1, factors = 3, rotation = "none")


Uniquenesses:(오차항의 분산:Psi)
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:(Lambda)
   Factor1 Factor2 Factor3
v1  0.808  -0.385   0.440
v2  0.752  -0.290   0.500
v3  0.813  -0.229  -0.530
v4  0.729  -0.139  -0.474
v5  0.802   0.521
v6  0.764   0.636


               Factor1 Factor2 Factor3
SS loadings      3.638   0.980   0.957
Proportion Var   0.606   0.163   0.159
Cumulative Var   0.606   0.770   0.929


#Goodness of Fit
The degrees of freedom for the model is 0 and the fit was 0.4755
```

```
f1<-factanal(x = m1, factors = 1, rotation = "none")

f2<-factanal(x = m1, factors = 2, rotation = "none")

f3<-factanal(x = m1, factors = 3, rotation = "none")

f4<-factanal(x = m1, factors = 4, rotation = "none")


trS<-sum(diag(cov(m1)))

loadings(f1)
```

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \boldsymbol{\epsilon} \ ,$$

$$\mathbf{x} = \begin{pmatrix} 2.22 \\ 2.44 \\ 2.22 \\ 2.38 \\ 2.22 \\ 2.38 \end{pmatrix} + \begin{pmatrix} 0.808 & -0.385 & 0.439 \\ 0.751 & -0.290 & 0.499 \\ 0.813 & -0.228 & -0.529 \\ 0.729 & -0.139 & -0.474 \\ 0.801 & 0.520 & 0.039 \\ 0.763 & 0.636 & 0.082 \end{pmatrix} \mathbf{f} + \begin{pmatrix} 0.005 \\ 0.101 \\ 0.005 \\ 0.224 \\ 0.084 \\ 0.005 \end{pmatrix}$$

3.2 잠재인자값(score)

```
# no rotation

> f1<-factanal(m1, factors=3, rotation="none", scores="regression")

> f1$scores
```

```
         Factor1     Factor2    Factor3
 [1,] -0.4849000 -0.61436654 -1.3867258
 [2,] -0.4727117 -0.62914545 -1.3548393
 [3,] -0.4796933 -0.61748582 -1.4000595
 [4,] -0.2404672  0.02470737 -1.2827153
...
[14,] -0.2297742 -0.30978440  1.2268066
[15,] -0.4742070 -0.94885832  1.1227961
[16,]  2.1435570  1.07880372 -0.4603225
[17,]  1.8969328  0.18998483  0.6897603
[18,]  2.4166552 -1.23605870 -0.1905737
```

3.3 인자회전(varimax:default)

```
factanal(m1, factors=3) # varimax is the default


Call:
factanal(x = m1, factors = 3)
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:
```

```
   Factor1 Factor2 Factor3
v1 0.944   0.182   0.267
v2 0.905   0.235   0.159
v3 0.236   0.210   0.946
v4 0.180   0.242   0.828
v5 0.242   0.881   0.286
v6 0.193   0.959   0.196

                Factor1 Factor2 Factor3
SS loadings       1.893   1.886   1.797
Proportion Var    0.316   0.314   0.300
Cumulative Var    0.316   0.630   0.929

The degrees of freedom for the model is 0 and the fit was 0.4755
```

3.4 인자회전(promax)

```
factanal(m1, factors=3, rotation="promax")


Call:
factanal(x = m1, factors = 3, rotation = "promax")
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:
   Factor1 Factor2 Factor3
v1          0.985
v2          0.951
v3                  1.003
v4                  0.867
v5  0.910
v6  1.033

               Factor1 Factor2 Factor3
SS loadings      1.903   1.876   1.772
Proportion Var   0.317   0.313   0.295
Cumulative Var   0.317   0.630   0.925

The degrees of freedom for the model is 0 and the fit was 0.4755
> # The following shows the g factor as PC1


> prcomp(m1)
Standard deviations:
```

```
[1] 3.0368683 1.6313757 1.5818857 0.6344131 0.3190765 0.2649086


Rotation:

         PC1          PC2        PC3        PC4        PC5          PC6

v1 0.4168038 -0.52292304  0.2354298 -0.2686501  0.5157193 -0.39907358

v2 0.3885610 -0.50887673  0.2985906  0.3060519 -0.5061522  0.38865228

v3 0.4182779  0.01521834 -0.5555132 -0.5686880 -0.4308467 -0.08474731

v4 0.3943646  0.02184360 -0.5986150  0.5922259  0.3558110  0.09124977

v5 0.4254013  0.47017231  0.2923345 -0.2789775  0.3060409  0.58397162

v6 0.4047824  0.49580764  0.3209708  0.2866938 -0.2682391 -0.57719858
```