다변량 분석 강의노트

Jinseog Kim
Department of Statistics and Information Science
Dongguk University
E-mail:jinseog.kim@gmail.com

November 26, 2007

Contents

1 Introduction

1.1 확률벡터

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = (x_1, x_2, \ldots, x_p)^T$$

기대값(평균벡터):

$$E\mathbf{x} = (Ex_1, Ex_2, \ldots, Ex_p)^T = (\mu_1, \mu_2, \ldots, \mu_p)^T$$

분산(공분산행렬, variance-covariance matrix:

$$\Sigma = V(\mathbf{x}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

여기서 $\sigma_{ii} = Var(x_i), i = 1, \ldots, p,$

$\sigma_{ij} = Cov(x_i, x_j), i \neq j$

So, using the vector-matrix form:

$$\Sigma = \mathrm{E}\left[(\mathbf{x} - \mathrm{E}[\mathbf{x}])\,(\mathbf{x} - \mathrm{E}[\mathbf{x}])^\top\right]$$

and

$$\mu = \mathrm{E}(\mathbf{x})$$

Note: The matrix Σ is "positive semi definite(양정치): For all non-zero vectors $\mathbf{z} \in C^p$,

$$\mathbf{z}^*\Sigma\mathbf{z} \geq 0$$

.

All eigenvalues λ_i of Σ are positive.

1.2 공분산행렬의 성질

Assume that

$\mathbf{x}, \mathbf{x_1}$ and $\mathbf{x_2}$ are a random $(p \times 1)$ vectors, $\mathbf{y}$ is a random $(q \times 1)$ vector, $\mathbf{a}$ is $(\mathbf{p} \times \mathbf{1})$ vector, A and B are $(p \times q)$ matrices.

- $\Sigma = \mathrm{E}(\mathbf{x}\mathbf{x}^\top) - \mu\mu^\top$

- $\mathbf{\Sigma}$ is positive-definite matrix (positive semi-definite)
- $\operatorname{var}(A\mathbf{x}+\mathbf{a}) = A\ \operatorname{var}(\mathbf{x})\, A^\top$
- $\operatorname{cov}(\mathbf{x},\mathbf{y}) = \operatorname{cov}(\mathbf{y},\mathbf{x})$
- $\operatorname{cov}(\mathbf{x_1}+\mathbf{x_2},\mathbf{y}) = \operatorname{cov}(\mathbf{x_1},\mathbf{y}) + \operatorname{cov}(\mathbf{x_2},\mathbf{y})$
- If $p = q$, then
  $\operatorname{var}(\mathbf{x}+\mathbf{y}) = \operatorname{var}(\mathbf{x}) + 2\operatorname{cov}(\mathbf{x},\mathbf{y}) + \operatorname{var}(\mathbf{y})$
- $\operatorname{cov}(A\mathbf{x}, B\mathbf{y}) = A\ \operatorname{cov}(\mathbf{x},\mathbf{y})\, B^\top$
- If $\mathbf{x}$ and $\mathbf{y}$ are independent, then

1.3 다변량 확률표본

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ 이 평균이 $\boldsymbol{\mu}_{1\times p}$이고 공분산 행렬이 $\Sigma_{p\times p}$인 다변량 분포에서 얻어진 확률표본이라고 하자. 여기서 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, 즉,

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{n1} \\ x_{12} & x_{22} & \ldots & x_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ x_{1p} & x_{2p} & \ldots & x_{np} \end{pmatrix} = X^\top$$

- 표본평균:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \frac{1}{n}X^\top\mathbf{1}_n = (\bar{x}_1, \ldots, \bar{x}_p)^\top, \bar{x}_k = \frac{1}{n}\sum_{i=1}^{n}x_{ki}, k = 1, \ldots, p.$$

- 표본공분산행렬:

$$S = (s_{kl}), s_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ki} - \bar{x}_k)(x_{li} - \bar{x}_l)$$

  or

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top,$$

  or

$$\frac{1}{n-1}X'(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}')X = \frac{1}{n-1}X'\left(I_n - \frac{1}{n}J\right)X.$$

다음과 같은 다변량 정규분포를 고려하자, 이 분포에서 20개의 난수를 발생시키고 발생된 난수에서 표본평균 및 표본 공분산 행렬을 구하라.

$$\boldsymbol{\mu} = (0,1,2)', \Sigma = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

```
library(MSBVAR)
## 다변량 정규분포에서 난수의 발생

## Calculate Covariance matrix
m<-c(0,1,2)
V<-matrix(c(3,2,1,2,4,1,1,1,5), ncol=3, nrow=3)

## generate errors from multivariate normal distribution
X<-rmultnorm(20, mu=m, vmat=V)
X<-t(X)

## 표본평균


## 표본분산
```

1.4 다변량 상관계수

확률벡터를 구성하는 변수들 사이의 관계를 측정하는 방법

- 상관계수(correlation coefficients)
- 다중상관계수(multiple correlation coefficients): 한 변수와 다른변수의 그룹간의 상관관계
- 부분상관계수(partial correlation coefficients):여러 개의 확률변수로 이루어진 두 개의 변수 그룹에서 한 그룹의 변수 값이 주어진 경우 다른 그룹내의 변수들간의 관계
- 정준상관계수(canonical correlation coefficients): 여러개의 확률변수로 이루어진 두 변수그룹간의 관계

1.5 상관계수(correlation coefficients)

$$R = D^{-1/2} S D^{-1/2}$$

여기서, $D = \begin{pmatrix} s_{11} & 0 & \dots & 0 \\ 0 & s_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_{pp} \end{pmatrix} = \text{diag}(s_{11}, s_{22}, \dots, s_{pp})$,이고, S는 공분산 행렬 (covariance matrix)이다.

1.6 다중상관계수(multiple correlation coefficients)

다변량 자료가 p개의 확률변수로 이루어져 있다고 가정하자.
p개 중에서 임의로 하나의 변수를 선택하고(이를 x_0)
나머지 $p-1$개에서 $q \le p-1$개를 선택하자(이를 $\mathbf{x}_{(q)}$로 표현하자). 여기서 x_0와 $\mathbf{x}_{(q)}$의 관계를 하나의 값으로 표현하기 위해서 x_0와 $\mathbf{x}_{(q)}$의 선형결합(linear combination), 즉

$$\boldsymbol{\beta}^\top \mathbf{x}_{(q)} = \beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_q x_{i_q},$$

로 표현하면 두 변수간의 상관계수를 구할 수 있다.

이렇게 모든 가능한 경우의 상관관계를 고려하고, 그 중에서 최대치를 다중상관계수라고 부른다.

2 행렬연산의 기초

2.1 행렬의 종류

- 정방행렬(square matrix) 행수와 열수가 같은 행렬 $A_{n\times n}$
- 대각행렬(diagonal matrix) 대각선원소를 제외한 모든 원소가 0인 행렬
- 항등행렬(Identity matrix): 대각행렬 중 대각선 원소가 모두 1인 행렬

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

- 정방 행렬의 원소가 모두 1인 행렬

$$J = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix} = \mathbf{1}\mathbf{1}'$$

- 전치행렬(Transpose matrix): 행렬 A의 원소 a_{ij}의 행번호와 열번호를 바꾼 행렬.

$$A' = A^\top = (a_{ji})$$

- 대칭행렬(Symmetric matrix):

$$A' = A.$$

- 역행렬(Inverse matrix): 행렬 A에 대하여, 어떤 행렬 B존재하여, 다음을 만족할 때, B를 역행렬이라고 하고 $B = A^{-1}$로 표현한다.

$$AB = BA = I.$$

- 직교행렬(Orthogonal matrix):

$$AA' = A'A = I, \text{혹은 } A' = A^{-1}.$$

2.2 선형변환

행렬 A, 열벡터 $\mathbf{b}$,

$$\mathbf{y} = A\mathbf{x} + \mathbf{b}$$

를 $\mathbf{x}$의 $\mathbf{y}$로의 선형변환(linear transformation)이라고 한다.

- If $E(\mathbf{x}) = \boldsymbol{\mu}$, $Cov(\mathbf{x}) = \Sigma_{\mathbf{x}}$, then

$$E(\mathbf{y}) = A\boldsymbol{\mu}_x + \mathbf{b},\ Cov(\mathbf{y}) = A\Sigma_{\mathbf{x}}A^{\top}.$$

- **Affine transformation**

  만일 A가 non-singular matrix(정칙행렬, 역행렬이 존재하는 경우)이면, $\mathbf{y} = A\mathbf{x} + \mathbf{b}$를 Affine transformation이라고 부른다.

2.3 선형독립(linearly independent)

- n개의 열벡터 $\mathbf{x}_1, \ldots, \mathbf{x}_n$가 0이 아닌 상수($a_i$)와의 선형결합

$$\mathbf{a}'\mathbf{x} = a_1\mathbf{x}_1 + \ldots + a_n\mathbf{x}_n = 0$$

  을 만족하면 $\mathbf{x}_1, \ldots, \mathbf{x}_n$를 선형종속(linearly dependent)라고 부르며, $\mathbf{a} = \mathbf{0}$이면 선형독립이라고 부른다.

- 행렬 X는 ,열벡터들 $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$으로 이루어 졌다고 할때, 선형독립인 열의 개수를 행렬 X의 계수(rank)라고 한다.

2.4 행렬식(determinant)

행렬식(determinant)는 정방 행렬 $A_{n\times n}$를 scalar값으로 mapping(대응)시킴. (det(A) 혹은 $|A|$로 표현)

의미: 행렬 A를 n차원 공간에서의 n개의 벡터라고 할 때, 그 벡터들로 만들어지는 도형의 부피(Volumn)의 개념으로 이해할 수 있다.

2.5 trace

정방행렬의 대각선 원소의 합을 행렬의 trace라고 한다.

$$tr(A + B) = tr(A) + tr(B)$$

$$tr(aA) = a \times tr(A)$$

$$tr(A) = tr(A')$$

.

If $A_{n\times n}$ matrix and $B_{n\times n}$ matrix, then

$$tr(AB) = tr(BA).$$

$$\mathrm{tr}(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}B_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n}B_{ji}A_{ij} = \sum_{j=1}^{n}(BA)_{jj} = \mathrm{tr}(BA)$$

$$tr(ABC) = tr(CAB) = tr(BCA)$$

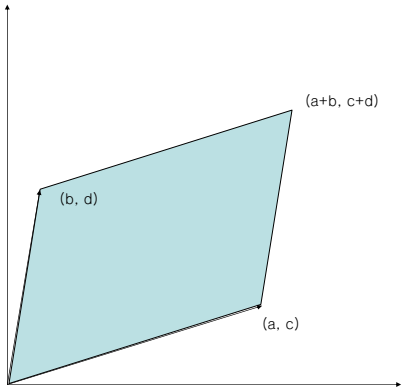


Figure 1: Graphical representation for the determinent

2.6 이차형식(Quadratic form)

$A_{p\times p}$ 가 대칭행렬(symmetric)이고 $\mathbf{x}$ 가 p 열벡터 일때,

$$Q(\mathbf{x}) = \mathbf{x}'A\mathbf{x}$$

를 $\mathbf{x}$의 A에 대한 이차형식이라고 부른다.

- 만일 $\mathbf{x}'A\mathbf{x} > 0$ (resp. $\mathbf{x}'A\mathbf{x} < 0$) for every vector $\mathbf{x} \neq 0$ 이면 A를 "양정치(positive definite)" (resp. 음정치(negative definite))행렬 이라고 부른다.
- If we change the strict inequality into $\geq; \leq$, A를 "semi-definite"(즉, 양반정치, 음반정치) 행렬 이라고 부른다.
- If $Q(v) < 0$ for some v and $Q(v) > 0$ for some other v, Q is said to be "indefinite".

이차형식은 χ^2 분포, mahalanobis 거리와 같은 데서 이용된다.

2.7 고유치, 고유벡터-eigen value and eigen vector

주어진 행렬 A 에 대하여 어떤 벡터($\mathbf{x}$)의 선형변환이 자기 자신($\mathbf{x}$)과 어떤 상수(λ)의 곱으로 표현할 수 있다고 하자. 즉,

$$A\mathbf{x} = \lambda\mathbf{x}.$$

이는

$$A\mathbf{x} = (\lambda I)\mathbf{x},$$

또는

$$(A - \lambda I)\mathbf{x} = 0$$

로 표현될 수 있다.

이 때, eigenvector는 위 식을 만족하는 $\mathbf{0}$이 아닌 벡터, eigenvalue는 역시 0이 아닌 스칼라 값이다. 다음의 예를 살펴보자.

$$A = \begin{bmatrix} 0 & 0 \\ -\frac{1}{2} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1-\lambda & 0 \\ -\frac{1}{2} & 1-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

여기서 eigen values와 eigen-vector는 $\lambda = 1, \mathbf{x} = \begin{bmatrix} 0 \\ c \end{bmatrix}$

See page 43–46 for the properties of eigen-value and eigen vectors.

2.8 멱등행렬(idempotent matrix)

$$A^2 = AA = A.$$

Example: 회귀분석에서 hat matrix $H = X(X'X)^{-1}X'$.

종속변수의 관측치를 행렬로 표현하면 $\mathbf{y}$이고, 회귀계수의 추정치는 $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$이다. 이 때, $\mathbf{y}$의 추정치는

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} = H\mathbf{y}.$$

또한 잔차벡터는

$$\mathbf{y} - X\hat{\boldsymbol{\beta}} = [I - X(X'X)^{-1}X']\mathbf{y} = (I - H)\mathbf{y}.$$

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = H.$$

이므로 $X(X'X)^{-1}X'$는 멱등행렬이다.

2.9 Spectral decomposition

Let A be a positive definite(pd) matrix. Then there exists a orthogonal matrix P, and diagonal matrix Λ,

$$A = P\Lambda P'.$$

If diagonal elements of Λ are the eigen values of A, then P is composed of the eigen-vectors, so

$$A = P\Lambda P'.$$

2.10 다변량자료의 산포 및 거리

다변량자료에서의 자료점들의 흩어져 있는 정도를 공분산행렬($S(\Sigma)$로 나타낸다. 이는 행렬로 표현되어 있어서 쉽게 이해하기가 어렵다. 따라서 하나의 스칼라값으로 표현하여 산포의 정도를 쉽게 나타낼 수 있다. 행렬의 행렬식(determinant)는 행렬에 포함되 있는 벡터들로 나타내어 지는 공간의 부피라고 한 적이 있다. 이를 공분산행렬에 적용하여 공분산행렬의 부피를 계산한 것이 일반화된 분산(generalized variance)이다.

$$|S| = \prod_{i=1}^{p} l_i$$

자료간의 거리는 ?

- Euclidian distance

$$\sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

- Mahalanovis distance

$$(\mathbf{x}-\mathbf{y})'S^{-1}(\mathbf{x}-\mathbf{y})$$

3 다변량정규분포

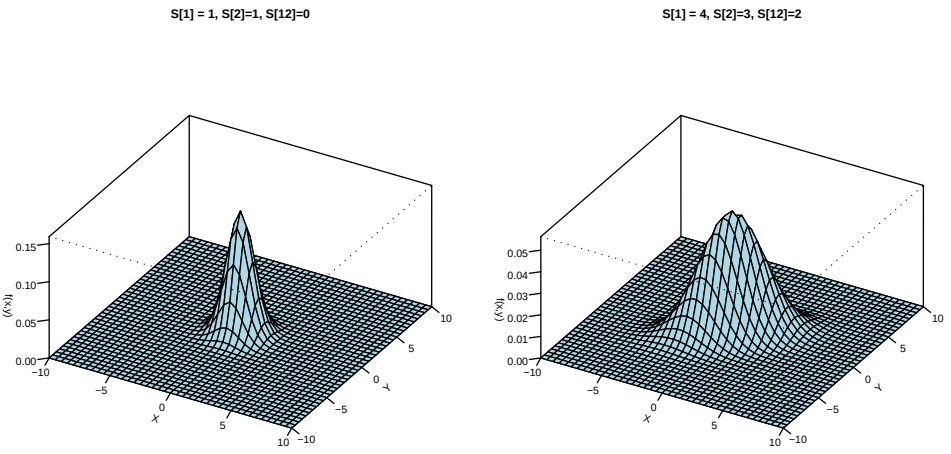
p 차원 확률벡터 $\mathbf{x} = (x_1, \ldots, x_p)'$가 다변량 정규분포를 따를 때,

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$$

로 표현하고 이에 대한 확률밀도함수는

$$f(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

```
>library(mvtnorm)

sigma <- matrix(c(4,2,2,3), ncol=2)
x<-seq(-10,10, by=0.5)

f<-function(x,y) { dmvnorm(c(x,y), mean=c(0,0), sigma=sigma)}
z <- matrix(ncol=length(x), nrow=length(y))
for(i in 1:length(x))
for(j in 1:length(x))
   z[i,j] <- f(x[i],x[j])

persp(x, y, z, theta = 30, phi = 30, expand = 0.5,
    col = "lightblue",
    ltheta = 120, ticktype = "detailed",
    xlab = "X", ylab = "Y", zlab = "f(x,y)",
    main="S[1] = 4, S[2]=3, S[12]=2")
```



3.1 다변량 정규분포의 표본분포

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ 이 서로 독립이고 $N_p(\boldsymbol{\mu}, \Sigma)$를 따를 때, 표본평균(벡터)은

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i,$$

표본분산(행렬)은

$$S = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

이다.

표본평균의 분포는

$$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \Sigma/n),$$

표본분산의 분포는

$$(n-1)S \sim W_p(n-1, \Sigma),$$

이며, 여기서 $W_p(n-1, \Sigma)$를 자유도가 $n-1$인 Wishart 분포라고 부른다.

일반적인 Wishart분포는 평균이 $\mathbf{0}$이고 분산이 Σ인 p차원 다변량정규분포에서 n개의 확률표본($\mathbf{z}_i, i = 1, \ldots, n$)을 추출할 때, 다음과 같은 확률행렬(random matrix)의 분포를 말한다.

$$\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \sim W_p(n, \Sigma).$$

$p = 1$이면 위의 분포는 카이제곱분포와 동일 하다. Recall that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

3.2 다변량정규분포의 성질

1. $\mathbf{x} \sim N_p(,) \iff \mathbf{a}'\mathbf{x} \sim N_1(,)$ for all $\mathbf{a} \neq 0$.

2. $\bar{\mathbf{x}}, S$는 서로 독립이다. (이 조건은 다변량 정규분포가 되기 위한 필요충분조건)

3. $\sigma_{ij} = 0, i \neq j \Longrightarrow x_i \perp x_j$

4. $A_{q\times p}$, $\mathbf{d}$ 상수벡터,

$$A\mathbf{x} + \mathbf{d} \sim N_q(A\boldsymbol{\mu} + d, A\Sigma A').$$

(예3.3):

5. $(\mathbf{x} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})' \sim \chi^2(p)$.

6. Assume

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then

(a)

$$\mathbf{x}_1 \perp \mathbf{x}_2 \Longleftrightarrow \Sigma_{12} = \mathbf{O}.$$

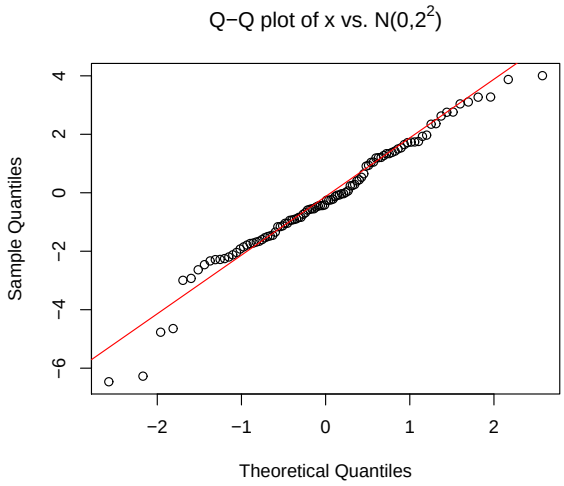
(b) Conditional distribution of $\mathbf{x}_1|_{\mathbf{x}_2} \sim N_q(\boldsymbol{\mu}_{1|2}, \Sigma_{11|2})$, where

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

3.3 Normality Test (정규성 검정)

- 일변량 QQ plot: 정규분포의 quantile과 자료의 quantile을 평면상의 점으로 표현하여 직선상에 위치하면 정규분포로 판단함

```
#qqplot
x<-rnorm(100, 0, 2)
qqnorm(x, main = expression("Q-Q plot of x vs. N(0," * 2^2 *")"))
qqline(x, col = "red")
```



- 카이제곱 그림: 표본의 Mahalanobis 거리(D_i^2)를 크기순으로 나열하고, 자유도 p인 카이제곱의 quantile과 비교(QQ plot), 여기서

$$D_i^2 \sim \chi^2(p), i = 1, \ldots, n.$$

  임을 이용한다, 단

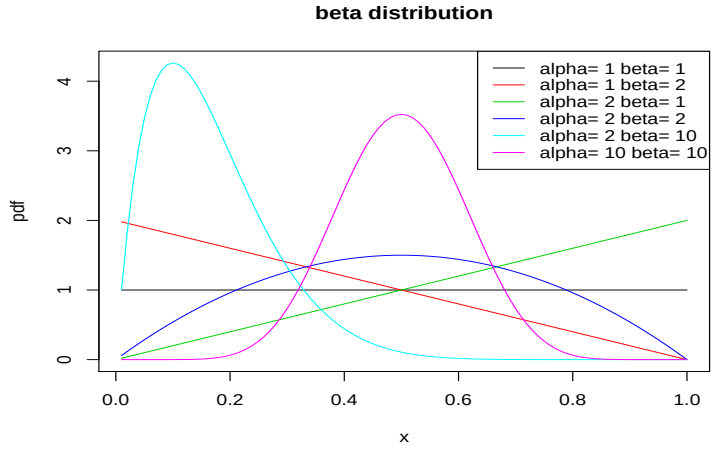
$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

- 베타 그림: 아래의 b_i 와 이에 대한 이론적 분포인 베타분포의 QQ plot.

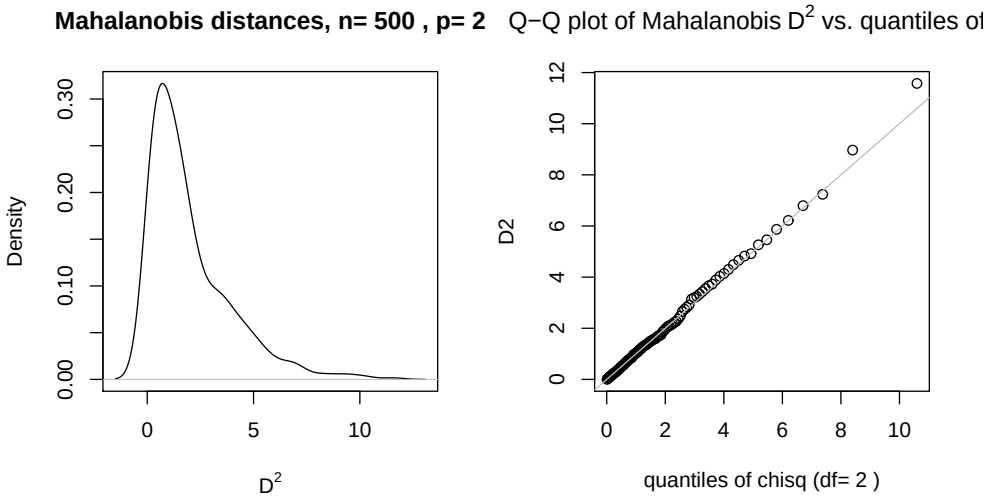
$$b_i = \frac{nD_i^2}{(n-1)^2} \sim Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$$

  (참고) Beta(α, β)분포의 pdf는

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1.$$

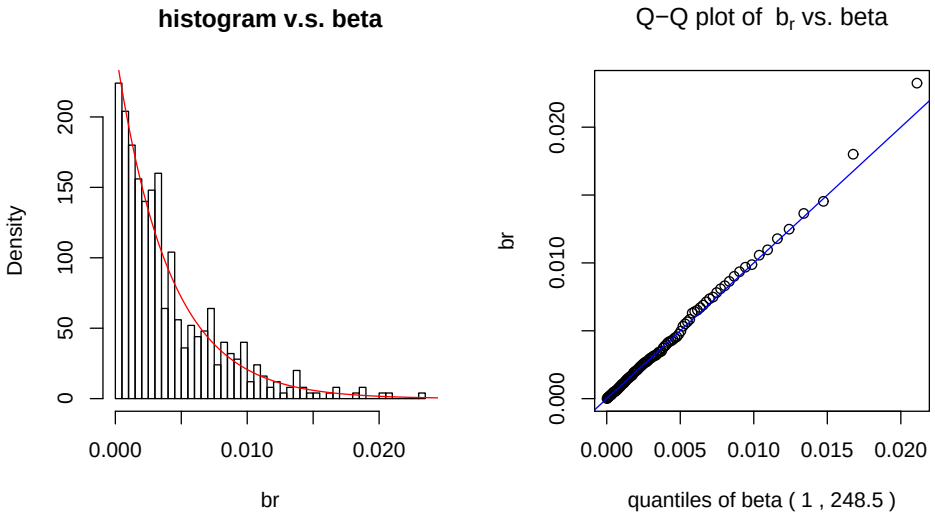


```
#chisq.plot
    n<-nrow(x)
    p<-ncol(x)
    Sx<-cov(x)
    mu<-apply(x, 2, mean)
    D2<-mahalanobis(x, center = mu, cov = Sx)
    par(mfrow=c(1,2))
    plot(density(D2, bw=.5), xlab=expression(D^2),
        main=paste("Mahalanobis distances, n=",n,", p=",p))
    qqplot(qchisq(ppoints(100), df=p), D2,
        xlab=paste("quantiles of chisq (df=",p,")"),
        main = expression("Q-Q plot of Mahalanobis" * ~D^2 *
                          " vs. quantiles of" * ~ chi^2))
    abline(0, 1, col = 'gray')
    par(mfrow=c(1,1))
```




```
#beta.plot
n<-nrow(x)
p<-ncol(x)
Sx<-cov(x)
mu<-apply(x, 2, mean) # or colMeans(x)
D2<-mahalanobis(x, center = mu, cov = Sx)
br<-n*D2/(n-1)^2
a<-p/2; b<-(n-p-1)/2 # beta분포의 모수
```

```
par(mfrow=c(1,2))
hist(br, nclass=50, freq=F, main="histogram v.s. beta")
lines(seq(0, 0.025, by=0.0001),
    dbeta(seq(0, 0.025, by=0.0001), a, b), col="red")
qqplot(qbeta(ppoints(100), shape1=a, shape2=b), br,
    xlab=paste("quantiles of beta (",a,",",b,")"),
    main = expression("Q-Q plot of " * ~b[r] * " vs. beta"))
abline(0, 1, col = "blue")
```



- 왜도(Skewness) and 첨도(Kurtoness) 일변량 확률변수의 왜도 및 첨도는

$$\beta_1 = E\left[\frac{x-\mu}{\sigma}\right]^3,$$

$$\beta_2 = E\left[\frac{x-\mu}{\sigma}\right]^4$$

로 정의되는데 이를 다변량으로 확장하면 아래와 같다 (Mardia, 1970).

$$\beta_{1p} = E\left[(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^3,$$

$$\beta_{2p} = E\left[(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^2.$$

만일 자료가 다변량 정규분포이면 $\beta_{1p} = 0$ 이고, $\beta_{2p} = p(p+2)$이 된다. 다변량 왜도 및 첨도를 구하기 위해서는 표본자료의 Mahalanobis 거리를 구하여 추정한다.

$$b_{1p} = \frac{1}{n^2}\sum_i\sum_j (D_{ij}^2)^3, b_{2p} = \frac{1}{n}\sum_i\sum_j (D_{ii}^2)^2.$$

3.4 다변량자료의 변수변환 (Box-Cox transformation)

다변량자료가 정규분포가 아닐 경우: 다변량 정규분포가 되도록 변수변환을 한다.

- Box-Cox변환 (Box and Cox, 1964): 모든 $i = 1, \ldots, n$과 $j = 1, \ldots, p$에 대하여

$$x_{ij}(\lambda_j) = \begin{cases} \frac{x_{ij}^{\lambda_j} - 1}{\lambda_j}, & \lambda_j \neq 0, \\ \log(x_{ij}), & \lambda_j = 0. \end{cases}$$

여기서 λ_j를 변환모수라고 부르며, 프로파일우도함수(Profile likelihood function)를 통해서 추정할 수 있다.

R에서는 `car` package에 Box-Cox transformation을 위한 함수 `box.cox.powers()` 가 있다.

```
library(car)
attach(Prestige)
box.cox.powers(cbind(Prestige$income, Prestige$education))
--
Box-Cox Transformations to Multinormality

     Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
[1,]    0.2617   0.1014        2.5799       -7.2799
[2,]    0.4242   0.4033        1.0517       -1.4278

L.R. test, all powers = 0:  7.694   df = 2   p = 0.0213
L.R. test, all powers = 1:  48.8727   df = 2   p = 0
Warning message:
NA/Inf replaced by maximum positive value in:
optimize(f = function(lambda) univ.neg.kernel.logL(x = X[, j],
--

par(mfrow=c(2,2))
qqnorm(Prestige$income, main="income")
qqnorm(Prestige$education, main="education")
qqnorm(Prestige$income^0.2617, main="Box-Cox Transform: income")
qqnorm(Prestige$education^0.4242, main="Box-Cox Transform: education")
```

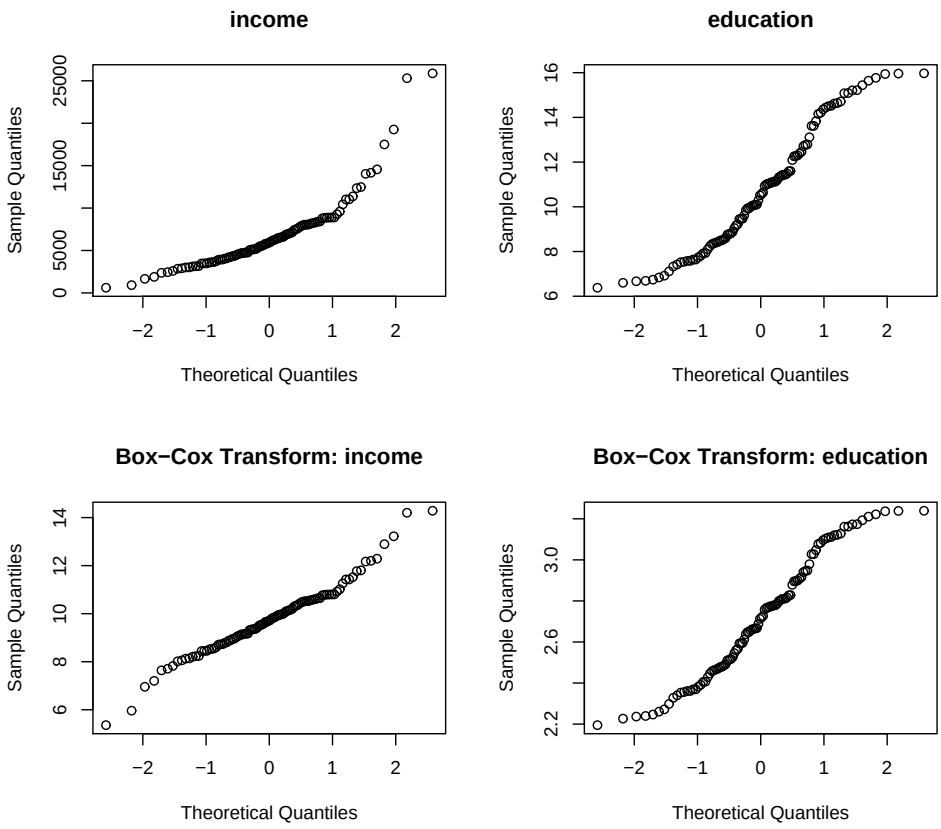


Figure 2: Box-Cox transformation

4 주성분분석 (Principal Component Analysis)

Page 151

주성분분석 (Principal Component Analysis)은 다차원자료를 설명력이 높은 몇개의 차원으로 축소하기 위한 분석방법이다.

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

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

인 선형변환을 고려하자. 이 때, 선형변환의 결과 $\mathbf{y}$의 공분산은

$$Cov(\mathbf{y}) = P'Cov(\mathbf{x})P = P'\Sigma P.$$

이 된다. 만일 $\mathbf{x}$의 선형변환 중에서 변환된 결과의 공분산행렬이 대각선원소를 제외하고 모두 0이면, 즉

$$Cov(\mathbf{y}) = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \lambda_p \end{pmatrix}$$

변환된 결과 $\mathbf{y} = (y_1, \ldots, y_p)'$ 는 모두 서로 독립이다. 또한 $Cov(\mathbf{y})$의 대각선 원소는 일변량 변수 $y_1, \ldots, y_p$ 각각에 대한 분산이 된다. 여기서 분산의 크기가 $Var(y_1) \geq \ldots Var(y_p)$라고 가정하자.

그러면 이러한 조건을 만족하는 선형변환 (P)은 어떻게 구할까?

Σ의 스펙트럴분해 (spectral decomposition)은

$$\Sigma = \Gamma\Lambda\Gamma'.$$

여기서 Γ 는 직교행렬(orthogonal)이고, Σ의 고유벡터로 구성되며, Λ는 대각행렬로써 대각선원소는 Σ의 고유치로 구성된다. 즉,

$$\Sigma = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \lambda_p \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_1' \\ \vdots \\ \boldsymbol{e}_p' \end{pmatrix}.$$

이런 사실을 이용하면

$$Cov(\mathbf{y}) = P'\Sigma P = P'\Gamma\Lambda\Gamma' P = \Lambda.$$

이기 위해서는

$$P = \Gamma$$

이면 된다. 이 때, $\mathbf{y}$를 $\mathbf{x}$의 주성분(Principal component)라고 부르며, 이 경우 공분산행렬의 대각원소의 합은 동일하다.

$$\sum_{j=1}^{p} Var(x_j) = \mathrm{tr}(\Sigma) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(y_j) \tag{1}$$

```
X<-cov(Prestige[,1:4]) #or var(...)
ei<-eigen(X)
LAMBDA<-diag(ei$values)
GAMMA<-ei$vectors
Z<-t(GAMMA)%*%X%*%GAMMA
diag(Z)
    [1] 1.802821e+07 8.287121e+02 1.298184e+02 1.816137e+00
sum(diag(Z))
    [1] 18029165
sum(diag(X))
    [1] 18029165
```

4.1 PCA 의 선택

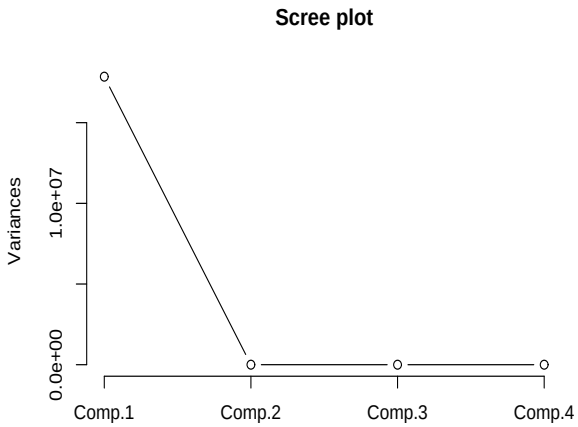
Scree Plot의 이용: 전체분산(변동) 중 주성분이 설명하는 변동의 양을 이용 전체 p개의 변수가 있을 때 다음을 계산하여 그림으로 표현해 준다.

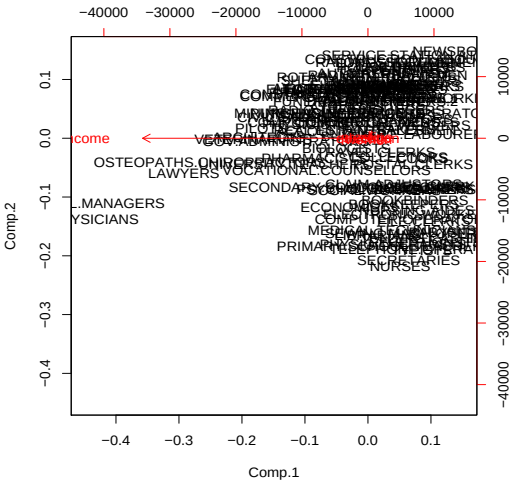
$$Var(y_1 + \ldots + y_q) = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, q = 1, \ldots, p.$$

4.2 R을 이용한 PCA

```
>pr1<-princomp(Prestige[,1:4])
>summary(pr1)
#--
Importance of components:
                          Comp.1  Comp.2     Comp.3    Comp.4
Standard deviation     4225.098 2.864e+01 1.133e+01 1.341e+00
Proportion of Variance    0.999 4.596e-05 7.200e-06 1.007e-07
Cumulative Proportion     0.999 9.999e-01 9.999e-01 1.000e+00
#--

>screeplot(pr1, type="lines", main="Scree plot")

>loadings(pr1) # Gamma: matrix of eigen vectors

Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4
education                0.126  0.991
```

```
income    -1.000
women             -0.987 -0.160
prestige          -0.156  0.979 -0.130


>biplot(pr1)
```





4.3 PCA의 활용

- 차원축소: 저차원을 이용한 Data Visualization
- 중회귀분석: 입력변수간의 다중공선성이 있을 때
- 요인분석, 판별분석, 집락분석, 이상치의 탐색 등.

4.4 실습 및 과제

- (과제) 교과서 p. 164 표8.3에 있는 자료에 대하여 주성분 분석을 하시오. (중간고사 전까지)
- (실습) MASS package 에 있는 UScrime data set을 이용하여 중회귀분석을 하시오.

5 Multidimensional scaling (MDS)

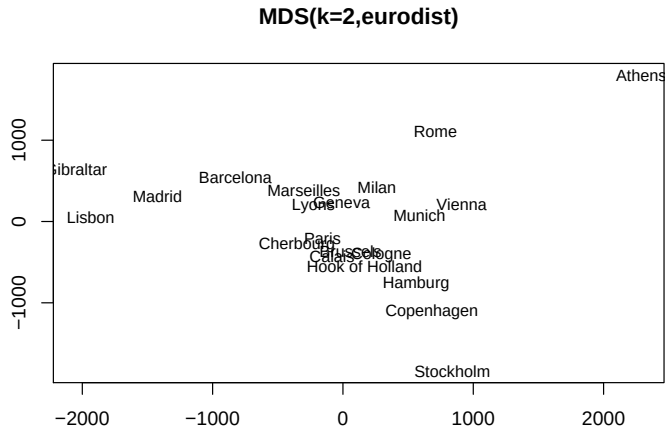
MDS는 R^p공간에 있는 자료점간의 유사도 (혹은 비유사도)를 측정하여 2차원 혹은 3차원 공간에 보여주는 통계방법론을 말한다. 즉

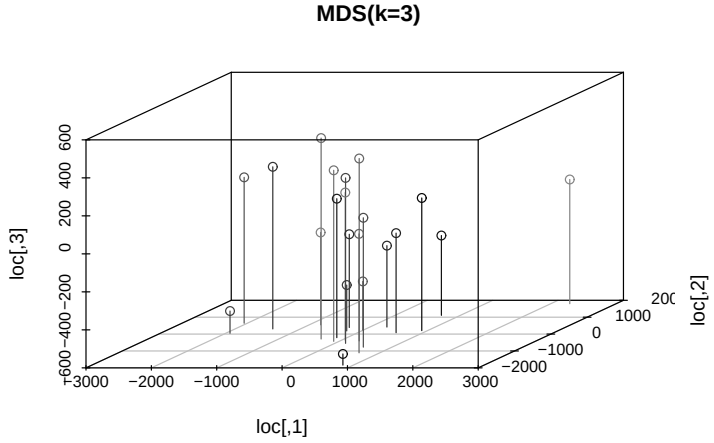
$$\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p \rightarrow \mathbf{z}_1, \ldots, \mathbf{z}_n \in R^k, \ k < p.$$

$\mathbf{x}_i$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))
```

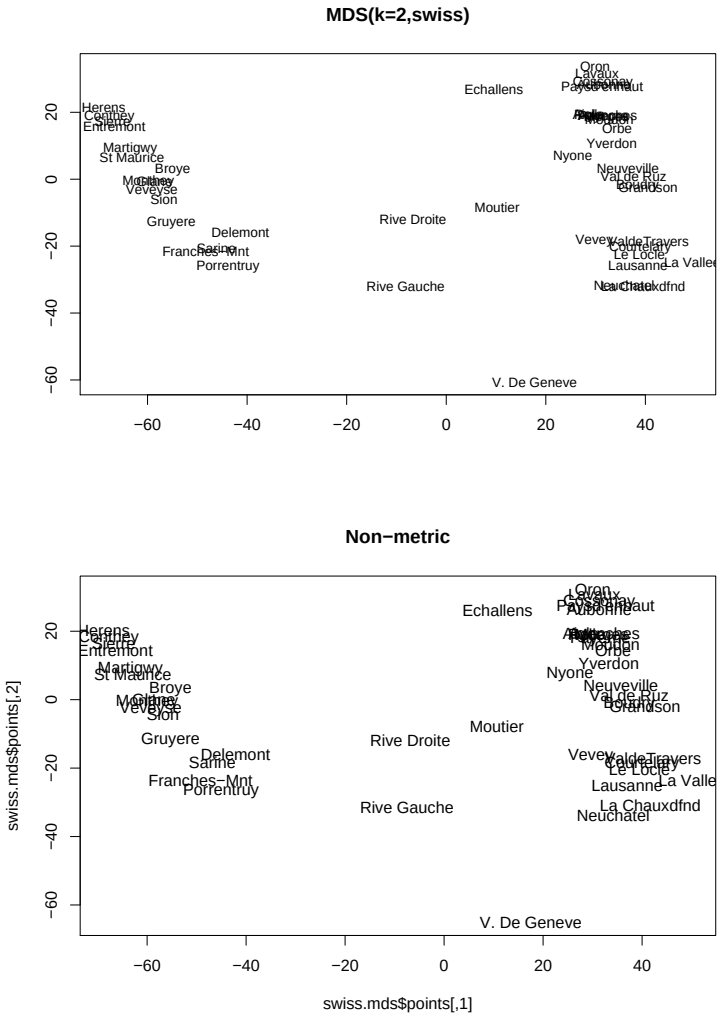
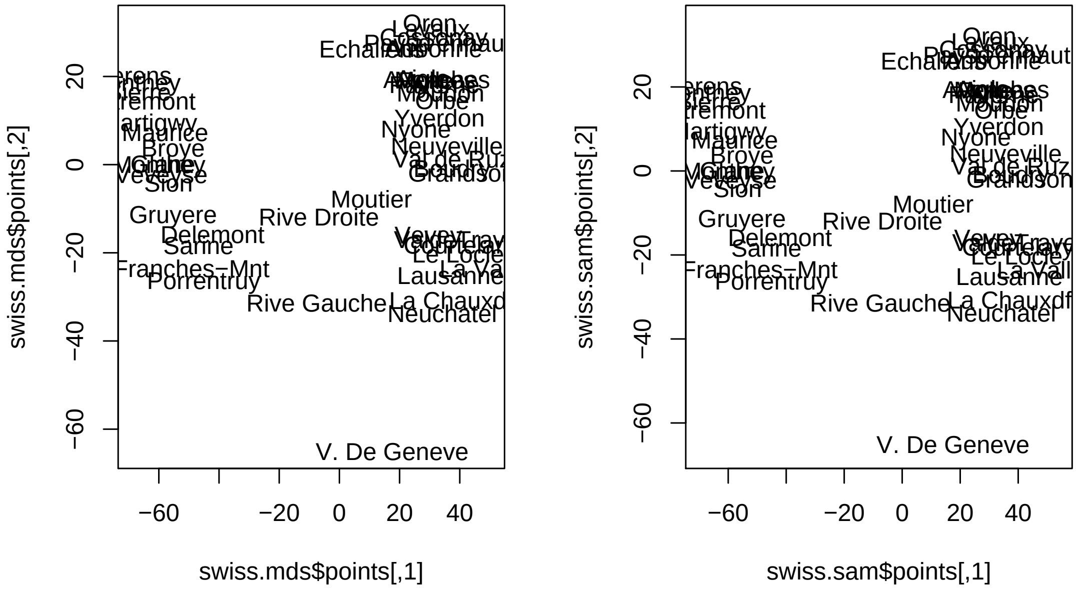


Figure 3: Metric and non-metric multi-dimensional scaling


swiss.mds$points[,2]
swiss.mds$points[,1]
−60
−20
0
20
40
Oron
Echallens
Orbe
Yverdon
Nyone
Neuveville
Val de Ruz
Grandson
Maurice
Broye
Sion
Moutier
Gruyere
Rive Droite
Delemont
Sarine
Vevey
Le Locle
Franches−Mnt
Porrentruy
Lausanne
Rive Gauche
La Chauxdf
Neuchatel
V. De Geneve
swiss.sam$points[,2]
swiss.sam$points[,1]

다변량자료분석 중간고사

1. (5pt) p 차원 랜덤벡터 $\mathbf{x}_1, \ldots, \mathbf{x}_n$이 서로 독립이고 $N_p(\boldsymbol{\mu}, \Sigma)$를 따를 때, 모평균 벡터($\boldsymbol{\mu}$) 및 모분산행렬($\Sigma$)의 추정량 및 추정량의 분포를 쓰시오.

2. (25pt) $\mathbf{x} = (x_1, \ldots, x_p)' \sim N_p(\boldsymbol{\mu}, \Sigma)$일 때 다음 중 다변량정규분포의 성질로 맞는 것은 O 틀리면 X 표하시오.

   (a) $\mathbf{x}$가 p차원 다변량 정규분포를 따르면 0이 아닌 모든 p 차원 상수벡터 $\mathbf{a}$에 대하여 $\mathbf{a}'\mathbf{x}$는 1변량 정규분포를 따른다.

   (b) 1번 문제에서 구한 추정량, 즉 표본 평균과 표본분산행렬은 서로 독립이다.

   (c) 다변량 정규분포에서 모분산행렬의 (i,j)번째 원소가 0 (즉 $\sigma_{ij} = 0, i \neq j$) 이면 확률변수 x_i와 x_j는 서로 독립이다.

   (d) $q \times p$행렬 $A_{q\times p}$와 p차원 상수벡터 $\mathbf{d}$에 대하여, $A\mathbf{x}+\mathbf{d}$는 평균이 $A\boldsymbol{\mu}+\mathbf{d}$ 이고 분산이 $A\Sigma A'+\mathbf{d}$인 다변량 정규분포를 따른다.

   (e) $(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})'$는 자유도가 p인 Wishart분포를 따른다.

3. (10pt) 다변량자료의 정규성 검정 (Normality Test) 방법들을 아는대로 나열하고 설명하시오.

4. (20pt) 고유치와 고유벡터를 설명하고 행렬 A 가 다음과 같이 주어져 있을 때, 고유치와 고유벡터를 구하는 R프로그램을 쓰고, 그 수행 결과를 해석하여라.

$$A = \begin{pmatrix} 1 & 3 & 4 & 1 \\ 2 & 9 & 2 & 1 \\ 2 & 5 & 2 & 0 \\ 5 & 5 & 3 & 6 \end{pmatrix}$$

5. (40pt) 아래자료를 다운로드한 후 주성분 분석을 하시오. 모든 R 프로그램을 쓰고, 결과는 요약하여 쓰시오

   `http://datamining.dongguk.ac.kr/lectures/Fall2007/multivariate/data.txt`

6 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 몇개의 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,

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

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.

6.1 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.

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

6.2 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.

6.3 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.

6.4 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.
- 평균고유값: 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 $\nu = \frac{1}{2}[(p-k)^2 - p - k]$

6.5 인자의 회전

해석이 용이한 단순한 구조로의 변환

- 직교회전(orthogonal roatation): 인자간에 서로 비상관이 되도록 회전

$$\Lambda_* = \Lambda G$$

  - Varimax: 인자적재값의 제곱의 분산을 최대화하여 인자적재값들이 서로 차이가 많이 나도록

$$\max \sum_{i=1}^{p} \sum_{j=1}^{k} (d_{ij}^2 - \bar{d}_j^2)^2, \quad d_{ij} = \lambda_{*ij} / \sum \lambda_{*ij}^2.$$

  - Quartimax:

- 비직교회전(non-orthogonal roatation) 또는 사각회전(oblique rotation)

  - Quatimin : 회전된 인자간의 공분산이 최소가 되도록
  - Covarmin
  - Promax

6.6 잠재인자값 (factor scores)의 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

6.7 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
```

6.7.1 No rotation

```
f1<-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 (분산비)

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

```
plot(f2$scores, main="Scatter plot of factor 1 and factor 2",
    xlab="factor 1",ylab="factor 2",pch=19,col="red")
text(f2$scores[,1], f2$scores[,2], 1:nrow(m1))
```

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

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

6.7.2 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
```

6.7.3 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
```

6.7.4 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
```

6.8 인자분석 절차 summary

6.9 실습 및 과제

- (실습) 교과서 p. 164 표8.3에 있는 자료에 대하여 요인 분석을 하시오.
- (실습) 교과서 p. 164 표8.3에 있는 자료에 대하여 요인 분석을 하시오.

7 판별분석-Discriminant analysis

- 몇개의 알려진 그룹으로부터 그 그룹을 구분해 주는 함수를 결정한다.
- 결정된(추정된) 판별함수를 이용하여 새로운 관측치를 분류한다.

7.1 Statistical Decision Theory

Let $\mathbf{X}_i, i = 1, \ldots, n$ be random vector and y_i be class labels (output variable).

여기서 입력변수 $\mathbf{X}$가 주어진 경우 y값을 예측하기 위한 함수 $f(\mathbf{X})$를 어떻게 찾을 수 있을까? 이를 해결하기 위하여 손실함수 (loss function)를 도입한다. 손실함수의 예로 제곱오차(squared error loss)를 고려하자.

$$L(Y, f(\mathbf{X})) = (Y - f(X))^2.$$

$f(\mathbf{X})$가 위의 손실함수에 대하여 모든 자료점들에 대하여 가장 작게되기 위해서는 다음의 예측오차(Expected squared prediction error) 를 최소화 하여야 한다.

$$EPE(f) = E(Y - f(X))^2 = \int (y - f(x))^2 Pr(dx, dy).$$

위의 예측오차(Expected squared prediction error)는 조건부 기대값을 이용하여 다시쓰면,

$$EPE(f) = E_X\{E_{Y|X}[(Y - f(X))^2|X]\}$$

이 된다. 따라서,

$$\arg\min_f EPE(f) = \arg\min_f E_X\{E_{Y|X}[(Y - f(X))^2|X]\} = E(Y|X = x),$$

즉, X가 주어질때 Y의 조건부 기대값이 된다.

만일, $y \in \{1, 2, \ldots, K\}$라고 하고, 손실함수를 0-1 loss를 고려한다면,

$$\begin{aligned}
\arg\min_f EPE(f) &= \arg\min_f E_X\{E_{Y|X}[1(Y \neq f(X))|X]\} \\
&= \arg\min_f E_X \left\{ \sum_{y \in \{1,2,\ldots,K\}} 1(Y \neq f(X)) P(Y = y|X) \right\} \\
&= \arg\min_f \sum_{y \in \{1,2,\ldots,K\}} 1(Y \neq f(X)) P(Y = y|X) \\
&= \arg\min_f (1 - P(Y = f(X)|X)) = \arg\max_f P(Y = f(X)|X)
\end{aligned}$$

위에서 마지막 식을 Bayes classifier라고 부른다. 그리고 Bayes classifier의 오분류률을 Bayes error라고 한다.

7.2 LDA and QDA

그룹을 $G \in \{1, 2, \ldots, K\}$라고 하고, 각 그룹에서 다변량 관측치가 얻어졌다고 하자. 이 때, 각 그룹 $k = 1, \ldots, K$ 에서 얻어진 관측치는 다변량분포를 따른다고 가정하자. 또한 관측치가 그룹 k에 속할 사전 확률은 π_k, $\sum_{k=1}^{K} \pi_k = 1$라고 하자.

여기서 $G = k$가 주어질 때, X의 조건부 확률을 $f_k(\mathbf{x})$라고 하면, Bayes 정리에 의하여 사후확률 분포 (posterior probability density)는

$$Pr(G = k|\mathbf{X} = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_k^K f_k(\mathbf{x})\pi_k}, \ k = 1, \ldots, K.$$

이다.

자료 $\mathbf{x}$가 어느 그룹에 속하는지를 판별한 때, 사후확률이 가장 큰 그룹에 할당하는 규칙을 Bayes rule이라고 한다. 즉,

$$\hat{G}(\mathbf{x}) = \arg\max_k \Pr(G = k|\mathbf{X} = \mathbf{x}) = \arg\max_k f_k(\mathbf{x})\pi_k. \tag{2}$$

만일 $\mathbf{x}$의 분포를 다변량 정규분포라고 하면, 즉,

$$f_k(\mathbf{x}) \sim MN_p(\boldsymbol{\mu}_k, \Sigma_k)$$

이때의 확률밀도함수는

$$f_k(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right).$$

이다.

어떤 $\mathbf{x}$가 두 분포 중 하나에서 나왔다고 가정하자 ($K = 2$). 위의 Bayes rule은 $\mathbf{x}$가 두 분포의 확률밀도함수를 비교하여 큰 값을 갖는 분포에서 나왔다고 간주할 수 있다. 즉,

$$\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x}) \Rightarrow \mathbf{x} \in group(1),$$

$$\pi_1 f_1(\mathbf{x}) < \pi_2 f_2(\mathbf{x}) \Rightarrow \mathbf{x} \in group(2).$$

식(2)를 최대화하기 위해서는 log변환된 값을 최대화 하는 것과 같다.

$$\max_k Q_k(\mathbf{x}) = \max_k \left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2}\ln|\Sigma_k| + \ln(\pi_k)\right\}$$

위의 괄호안의 함수는 이차함수가 되므로 이를 이차판별함수라고 부르고, 이차판별함수를 이용한 판별분석을 이차판별분석(QDA)라고 부른다.

만일 $\Sigma_1 = \ldots = \Sigma_K = \Sigma$라고 하면, 즉, 다변량 확률밀도함수가 아래와 같다고 하면,

$$f_k(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_k)\right), k = 1, \ldots, K$$

$$\begin{aligned}
\max_k Q_k(\mathbf{x}) &= \max_l \left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \ln(\pi_k)\right\} \\
&= \max_k \left\{\mathbf{x}^T\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}_k + \ln(\pi_k)\right\}
\end{aligned}$$

즉 1차함수를 최대화하는 k를 찾으면 된다. 이 때 다음함수를 선형판별함수라고 부른다.

$$L_k(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}_k + \ln(\pi_k)$$

7.3 R에서의 판별분석

R은 선형판별분석과 이차판별분석을 위하여 lda()와 qda() 두 가지 함수를 제공하고 있다. 이 함수들은 MASS 라이브러리안에 내장되어 있다.

```
library(MASS)
#Use the iris data
#Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
train <- sample(1:150, 75)
l1<-lda(Species~., iris[train,])



Prior probabilities of groups:
    setosa versicolor  virginica
 0.3866667  0.2800000  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.955172    3.389655     1.472414   0.2448276
versicolor     5.947619    2.780952     4.247619   1.3476190
virginica      6.564000    2.908000     5.596000   2.0400000

Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  0.9949202 -0.1047866
Sepal.Width   1.2351616  2.1893276
Petal.Length -2.5231096 -0.5254605
Petal.Width  -2.2256767  1.9989983

Proportion of trace:
   LD1    LD2
0.9962 0.0038
```

8 Cluster analysis (군집분석)

- Here we observe the features $X_1, X_2, \ldots, X_p$, but no outcome measure Y. Denote the data by $\mathbf{x}_i; i = 1, 2, \ldots, N$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
- Objectives of cluster analysis is that the observations set into natural clusters (groups) in which member is as similar to each other and as different as possible from the members of the other groups.

8.1 Applications: 응용사례

Clustering algorithms have been used in a large variety of applications [Jain and Dubes 1988; Rasmussen 1992; Oehler and Gray 1995; Fisher et al. 1993]. We describe several applications where clustering has been employed as an essential step. These areas are: (1) image segmentation, (2) object and character recognition, (3) document retrieval, and (4) data mining.

- Clustering for classification- handwritten zip code problem- can we find prototype digits for 1,2, etc, to use for classification

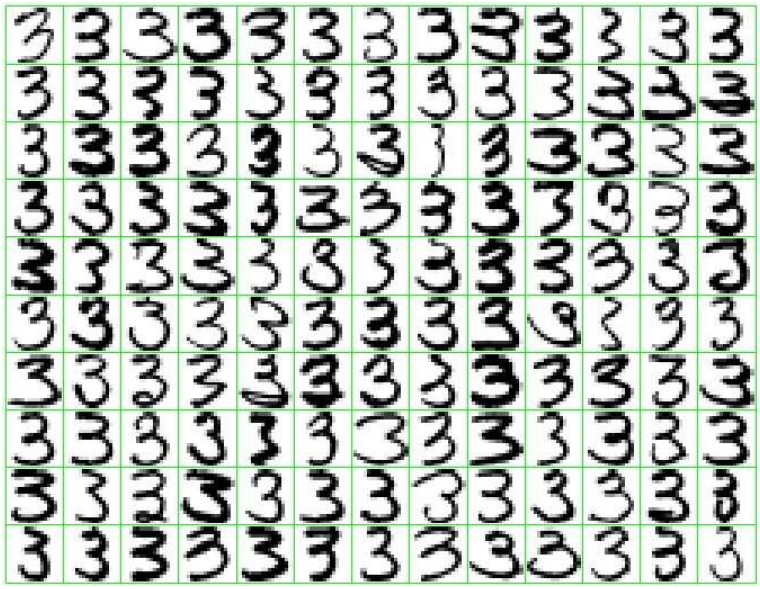


Figure 4: handwritten zip code problem

- DNA Microarray data - which samples cluster together? Which genes cluster together?
- Image Segmentation- a fundamental component in many computer vision

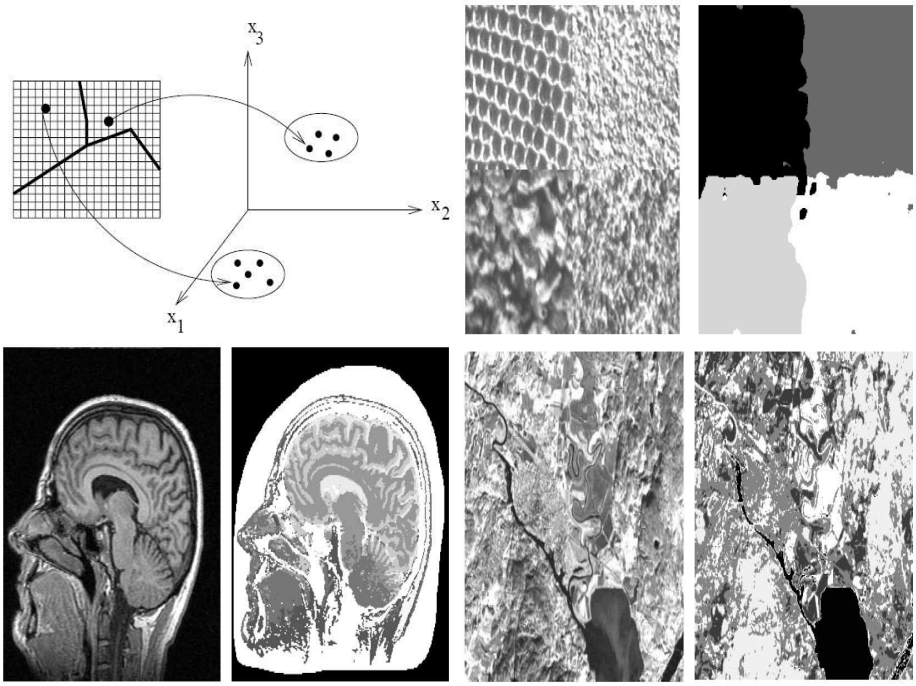


Figure 5: Image segmentation and texture example and MRI image

8.2 R libraries related with clustering

- `stat` hclust
- `cluster` Cluster Analysis Extended Rousseeuw et al.
- `clusterfly` Explore clustering interactively using R and GGobi
- `clustvarsel` Variable Selection for Model-Based Clustering
- `gclus` Clustering Graphics
- `mclust` Model-Based Clustering / Normal Mixture Modeling

Table 1: R functions

Function	Description	Package
hclust	Hierarchical Clustering	stats
kmeans	K-Means Clustering	stats
dist	distance matrix computed by using the specified distance measure	stats

8.3 Dissimilarity Measures (비유사도)

Clustering은 data point들간의 미리 정해진 거리를 측정하여 이들을 몇 개의 서로 다른 group으로 분류하는 것이다.

When all features are continuous:

- **Minkowski metric**

$$\begin{aligned} d_p(\mathbf{x}_i, \mathbf{x}_j) &= \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^p \right)^{1/p} \\ &= ||\mathbf{x}_i - \mathbf{x}_j||_p. \end{aligned}$$

$$\begin{cases} \sum_{k=1}^{d} |x_{i,k} - x_{j,k}| = ||\mathbf{x}_i - \mathbf{x}_j||_1, & \text{Manhatan distance } (p=1) \\ \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2 \right)^{1/2} = ||\mathbf{x}_i - \mathbf{x}_j||_2 & \text{Euclidean distance } (p=2) \\ ||\mathbf{x}_i - \mathbf{x}_j||_\infty = \max_{k=1,\ldots,d} |x_{i,k} - x_{j,k}| & \text{Supremum norm} (p=\infty). \end{cases}$$

- Mahalanobis distance: When the linear correlations exist among features,

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)'.$$

Categorical variables:

- **Hamming distance**

$$d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{d} I(x_{i,k} \neq x_{j,k})$$

if $\mathbf{x}_1 = (1,0,1,1,1,0,1)'$ and $\mathbf{x}_2 = (1,0,0,1,0,0,1)$ then

$$d_H(\mathbf{x}_1, \mathbf{x}_2) = 2.$$

- **binary distance**

$$d_B(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^{d} \{I(x_{i,k} = 1 \text{ and } x_{j,k} = 0) + I(x_{i,k} = 0 \text{ and } x_{j,k} = 1)\}}{\sum_{k=1}^{d} I(x_{i,k} = 1 \text{ or } x_{j,k} = 1)}$$

if $\mathbf{x}_1 = (1,0,1,1,1,0,1)'$ and $\mathbf{x}_2 = (1,0,0,1,0,0,1)$ then

$$d_B(\mathbf{x}_1, \mathbf{x}_2) = 2/5.$$

```
x <- matrix(rnorm(100), nrow=5)
dist(x)
dist(x, method= "manhatan")
dist(x, method= "maximum")

x <- c(0, 0, 1, 1, 1, 1)
y <- c(1, 0, 1, 1, 0, 1)
dist(rbind(x,y), method= "binary")
    x
y 0.4

hamming<-function(x,y){sum(x != y)}

hamming(x,y)
[1] 2
```

8.4 Types of Clustering methods

- Hierarchical-clustering
- Non-Hierarchical-clustering or Partition based clustering

8.5 Hierarchical-clustering (계층적 군집분석)

N개의 관측값들이 있을 때,

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N.$$

각 자료들간의 거리를 계산한다.

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$$

- 가장 가까운 관측값을 하나의 그룹으로 만든다.
- $N-1$개의 군집에서 다시 가장 가까운 두 그룹을 묶는다.
- 이런 방식으로 모든 관측값이 하나의 군집이 될 때까지 반복한다.

8.5.1 두 군집간의 거리

```
> as.matrix(dist(x))
         1         2         3         4         5
```

```
1 0.000000 6.849552 6.126787 7.187622 5.476626
2 6.849552 0.000000 7.577749 6.016812 6.645836
3 6.126787 7.577749 0.000000 7.220424 6.165865
4 7.187622 6.016812 7.220424 0.000000 6.061572
5 5.476626 6.645836 6.165865 6.061572 0.000000
```

- 최단연결법(Single linkage): 두 군집간 자료들간의 거리 중에서 가장 가까운 거리
- 최장연결법(Complete linkage):두 군집간 자료들간의 거리 중에서 가장 먼 거리
- 평균연결법(Average linkage): 두 군집간 자료들간의 거리의 평균
- 중심연결법(Centroid linkage):두 군집의 중심(평균)간의 거리
- 중앙값연결법(Median linkage): 두 군집의 중심을 군집내 자료들의 중앙값으로 정의

최단거리 연결법을 이용하면, 위의 거리 행렬에서는 1과 5를 맨 처음 하나의 그룹으로 묶는다.

```
        (1,5)          2        3        4
(1,5) 0.000000   6.645836 6.126787 6.061572
   2             0.000000 7.577749 6.016812
   3                      0.000000 7.220424
```

그 다음 2과 4를 다시 하나의 그룹으로 묶는다.

```
        (1,5)       (2,4)        3
(1,5) 0.000000   6.061572 6.126787
(2,4)            0.000000 7.220424
```

다음 (1,5)와 (2,4)의 거리가 가장 작으므로 이를 하나의 그룹으로 묶는다.

```
                  3
(1,5,2,4)   6.126787
```

```
par(mfrow=c(2,2))
plot(h<-hclust(dist(x), method = "single"))
plot(h<-hclust(dist(x), method = "complete"))
plot(h<-hclust(dist(x), method = "average"))
plot(h<-hclust(dist(x), method = "centroid"),hang=-1)
```

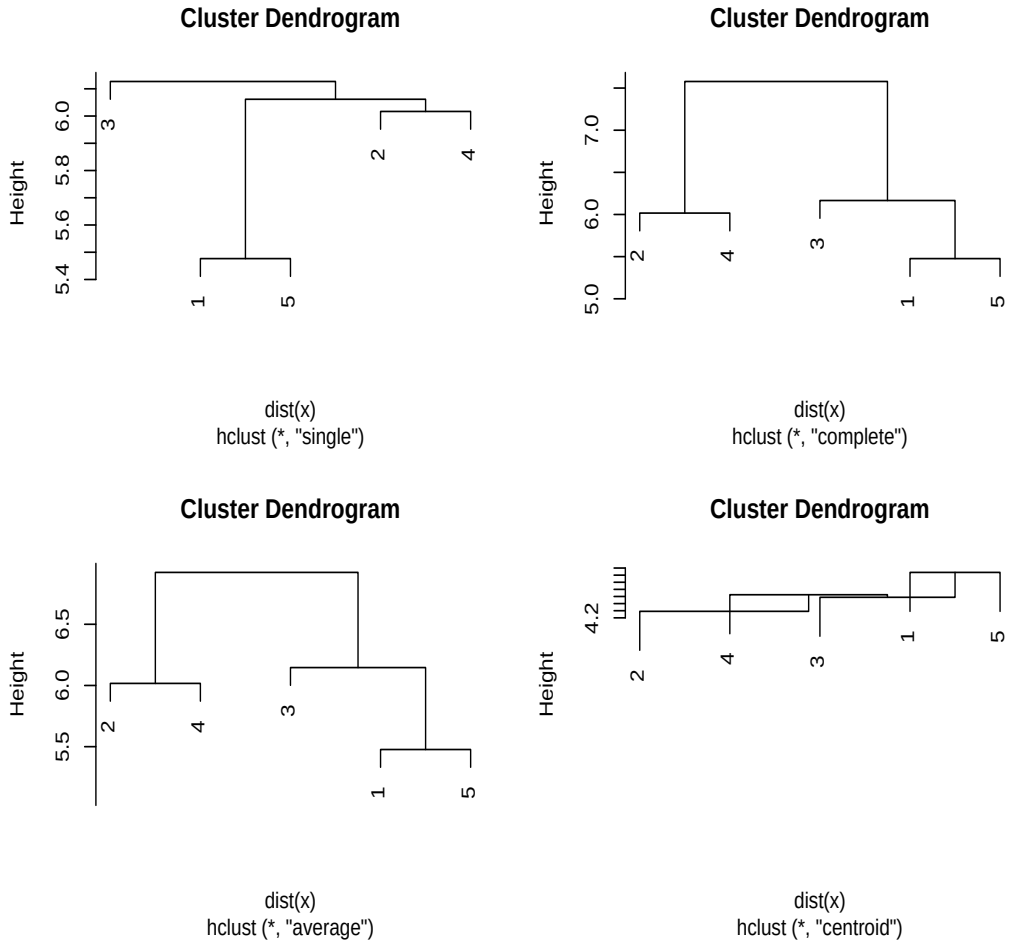


Figure 6: 군집간 연결방법에 따른 Dendrograms

8.6 k-means

1. For each data point, the closest cluster center (in Euclidean distance) is identified;
2. Each cluster center is replaced by the coordinatewise average of all data points that are closest to it.
3. Steps 1 and 2 are alternated until convergence. Algorithm converges to a local minimum of the within-cluster sum of squares.
4. Typically one uses multiple runs from random starting guesses, and chooses the solution with lowest within cluster sum of squares.

```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))
```

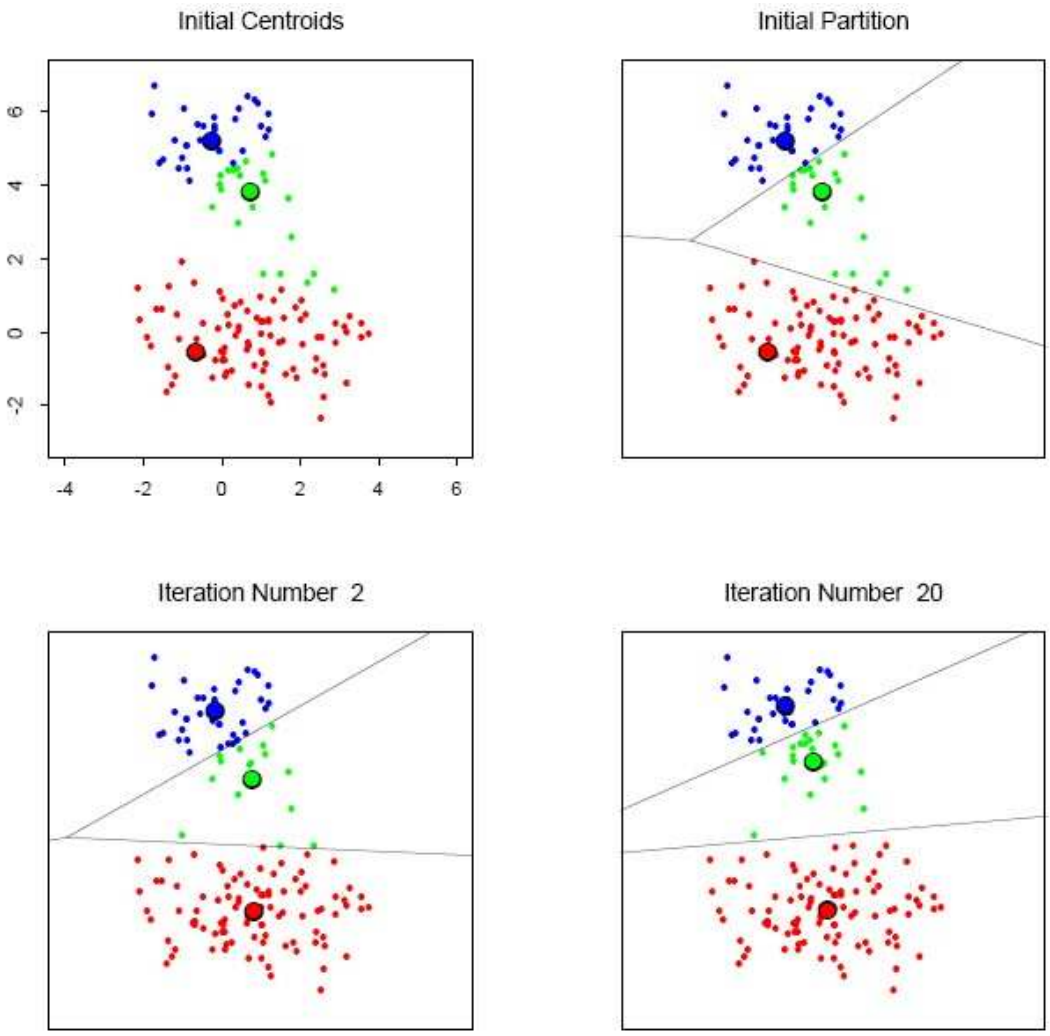


Figure 7: Successive iterations of the K-means clustering algorithm for the simulated data

INPUT ARGUMENTS:

x :data matrix

centers : cluster의 개수 혹은 cluster 수에 해당되는 초기값 벡터들.

iter.max: 최대 반복수

nstart: If centers is a number, how many random sets should be chosen?

algorithm:

결과:

$cluster: A vector of integers indicating the cluster

$centers: A matrix of cluster centres.

$withinss: The within-cluster sum of squares for each cluster.

```
$size: The number of points in each cluster
```

```
clusplot(x, rbinom(25,2, 0.5)+1, shade=F, color=T, lines=0)
```

8.7 군집수의 결정

Define a measure of "quality" of the partition in K clusters: Using so-called internal indexes, e.g. dissimilarity/distance within the clusters Based on the values of this measure on $K = (1), 2, \ldots$ use a rule to chose K:

Squared distances from centroids (within clusters sum of squares) is

$$W(K) = \sum_{j=1,\ldots,K} \sum_{i \in C_j} d^2(x_i, \bar{x}_j).$$

It takes a low values when the partition is good. However, it is monotone non-increasing in K. Consider the following measure,

- *Calinski and Harabasz (1974)* choose K to maximize

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)},$$

  where $B(K)$ is the sum of squares between cluster means.

- *Hartingan (1975)* selects the smallest K if $H(K) \leq 10$ where

$$H(K) = \frac{W(K) - W(K+1)}{W(K+1)} \Big/ (n - K - 1).$$

- *Krzanowski and Lai (1985)*: choose K to maximize

$$KL(K) = \left| \frac{DIFF(K)}{DIFF(K+1)} \right|,$$

  where $DIFF(K) = (K-1)^{2/p} W(K-1) - K^{2/p} W(K)$, and p is the dimension of variables.