다변량 분석

Jinseog Kim

October 2, 2007

1 R의 기초

1.1 R의 시작

- Download the R software from
  - http://www.r-project.org/ or
  - http://bibs.snu.ac.kr/R/.
- Install R software (R-2.5.0-win32.exe: Setup program (about 29 megabytes)).

1.2 R objects

```
vector,        matrix,      list
data.frame,    factor,      function
...
```

1.3 Indexing

R has three indexing constructs:

```
object [ arg1, ... , argn ]     # for vector, matrix, array
object [[ arg1, ... , argn ]]   # for list
object $ tag                    # for data.frame or named list
```

1.4 R operators

```
-   :Minus, can be unary or binary
+   :Plus, can be unary or binary
*   :Multiplication, binary
/   :Division, binary
%%  :Modulus, binary
<   :Less than, binary
>   :Greater than, binary
```

```
==   :Equal to, binary
>=   :Greater than or equal to, binary
<=   :Less than or equal to, binary
!    :Unary not
:    :Sequence, binary (in model formulae: interaction)
^    :Exponentiation, binary
&    :And, binary, vectorized &&  :And, binary, not vectorized
|    :Or, binary, vectorized ||  :Or, binary, not vectorized
<-   :Left assignment, binary,
<<- : global assignment
->   :Right assignment, binary
=    :Left ssignment, binary
$    :List subset,  binary
```

1.4.1 R operators:Example

%*% :Matrix product, binary

```
> a
     [,1] [,2]
[1,]    3    6
[2,]    4    8
> b
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> a%*%b
     [,1] [,2]
[1,]   15   33
[2,]   20   44
```

1.4.2 R operators:Example

```
%o% :Outer product, binary
> c(1,2) %o% c(3,4)
     [,1] [,2]
[1,]    3    4
[2,]    6    8

%x% :Kronecker product, binary

> a                    > b                    > b%x%a
      [,1] [,2]            [,1] [,2]              [,1] [,2] [,3] [,4]
[1,]    3    6       [1,]    1    0       [1,]    3    6    0    0
[2,]    4    8       [2,]    0    1       [2,]    4    8    0    0
                                          [3,]    0    0    3    6
                                          [4,]    0    0    4    8
```

1.5 Flow control

```
if ( cond ) expr
if ( cond ) expr1 else expr2
while ( cond ) expr
repeat expr
for ( var in list ) expr
```

Within the loop constructs (while, repeat, for), one may use

- break (to terminate the loop) and
- next (to skip to the next iteration).

1.6 R Functions: two simple examples

```
> name <- function(arg_1, arg_2, ...) expression
```

(예제)mile을 km로 바꾸는 프로그램.

```
miles.to.km <- function(miles) miles*8/5
```

```
> miles.to.km(175) # Approximate distance to Sydney, in miles
[1] 280
```

```
만일 100, 200 300 miles를 kilometer로 바꾼다면
> miles.to.km(c(100,200,300))
[1] 160 320 480
```

1.7 Common Useful Functions

```
print()     # Prints a single R object
cat()       # Prints multiple objects, one after the other
length()    # Number of elements in a vector or of a list
mean()
median()
```

```
range()
sum()
unique()    #Gives the vector of distinct values
diff()      # Replace a vector by the vector of first differences
            # diff(x) has one less element than x
sort()      # Sort elements into order, but omitting NAs
order()     # x[order(x)] orders elements of x, with NAs last
rev()       # reverse the order of vector elements
cumsum()
cumprod()
```

1.8 평균,표준편차를 구하는 프로그램

```
mean.and.sd <- function(x=1:10) {
 av <- mean(x)
 sd <- sqrt(var(x))
```

```
 c(mean=av, SD=sd)
}


> mean.and.sd()
    mean      SD
5.500000 3.027650
```

1.9 Reading data from files

If there is a text file named houses.data as follows:

```
      Price    Floor     Area    Rooms      Age  Cent.heat
01    52.00    111.0      830        5      6.2         no
02    54.75    128.0      710        5      7.5         no
03    57.50    101.0     1000        5      4.2         no
04    57.50    131.0      690        6      8.8         no
```

```
05    59.75      93.0       900      5         1.9      yes
...
```

The data frame may then be read as

```
> HousePrice <- read.table("houses.data", header=TRUE)
```

1.10 다변량정규분포

```
\library(MASS)
Sigma <- matrix(c(10,3,3,2),2,2)
Sigma
x<-mvrnorm(n=1000, rep(0, 2), Sigma)
var(X)
```

2 Introduction

2.1 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:

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

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.

2.2 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.

- $\Sigma = \mathrm{E}(\mathbf{x}\mathbf{x}^\top) - \mu\mu^\top$
- $\boldsymbol{\Sigma}$ is positive-definite matrix (positive semi-definite)
- $\mathrm{var}(A\mathbf{x} + \mathbf{a}) = A\ \mathrm{var}(\mathbf{x})\ A^\top$

- $\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \operatorname{cov}(\mathbf{y}, \mathbf{x})$
- $\operatorname{cov}(\mathbf{x_1} + \mathbf{x_2}, \mathbf{y}) = \operatorname{cov}(\mathbf{x_1}, \mathbf{y}) + \operatorname{cov}(\mathbf{x_2}, \mathbf{y})$
- 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

2.3 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,

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

- 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
```

2.4 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

2.5 상관계수(correlation coefficients)

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

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

2.6 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 선형결

합(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.

3 행렬연산의 기초

3.1 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):

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

3.2 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.

3.3 선형독립(linearly independent)

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

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

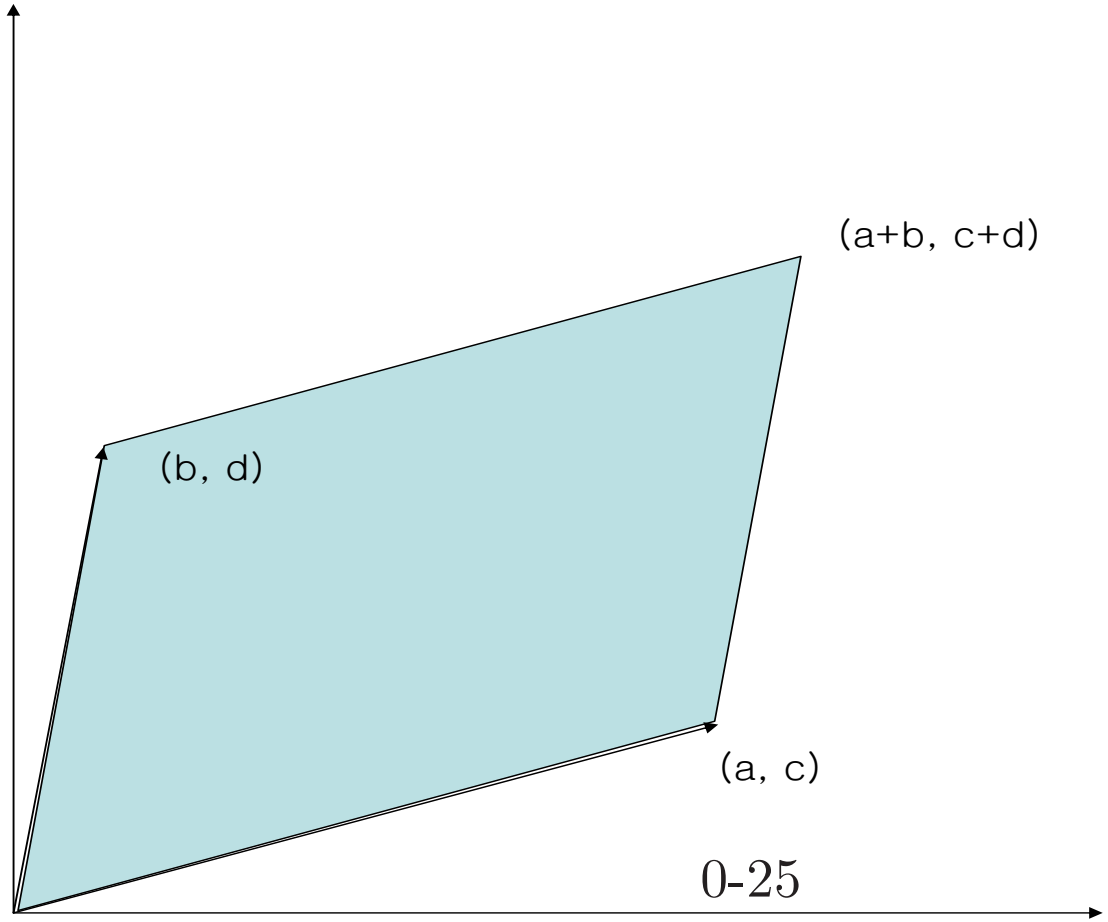
  을 만족하면 $\mathbf{x}_1, \ldots, \mathbf{x}_n$를 선형종속(linearly dependent)라고 부르며, $\mathbf{a} = \mathbf{0}$이면 선형독립이라고 부른다.

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

3.4 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.


(a+b, c+d)
(b, d)
(a, c)
0-25

3.5 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

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

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

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

3.6 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.

3.7 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.

3.8 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.

3.9 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

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

3.10 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

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

4 다변량정규분포

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

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

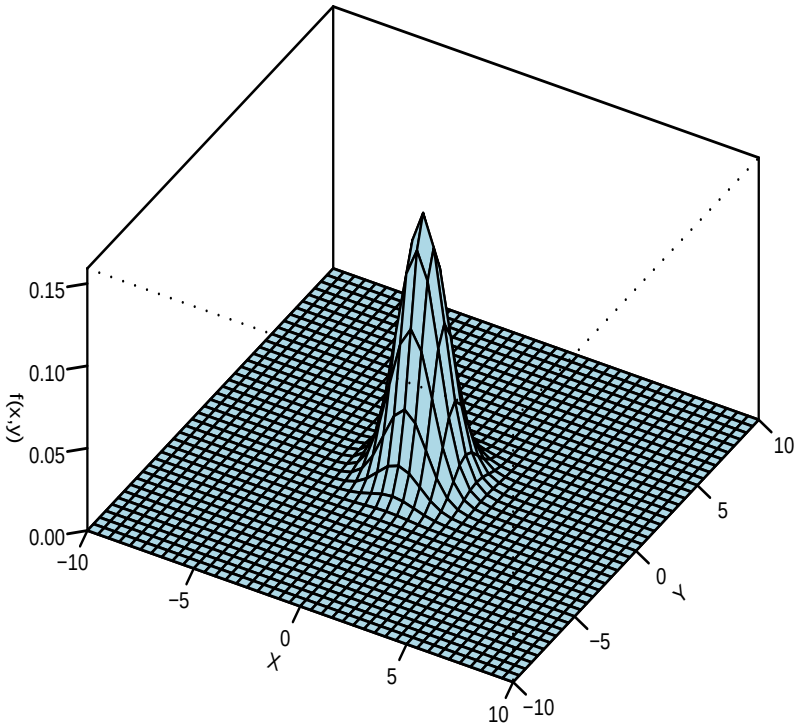
로 표현하고 이에 대한 확률밀도함수는

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

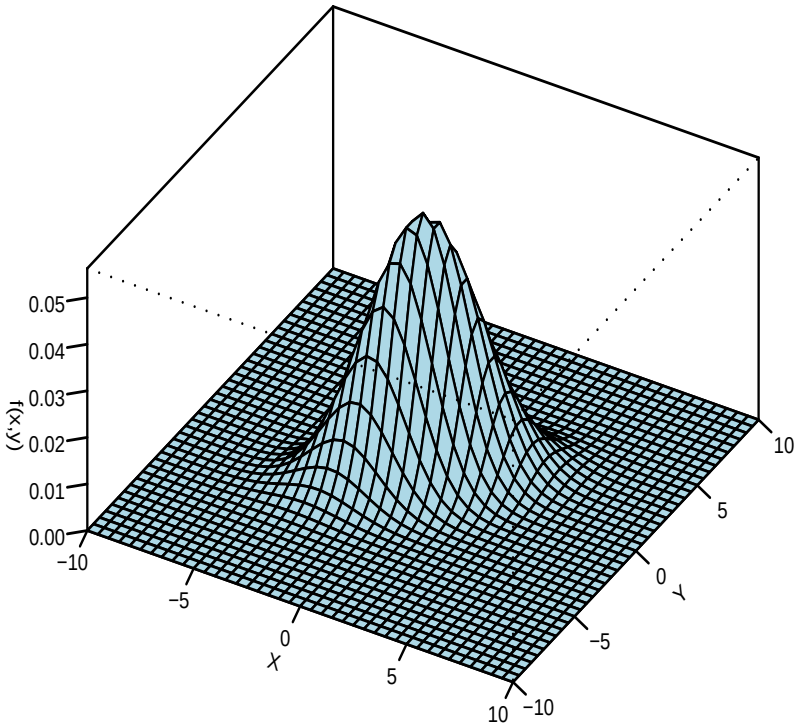
```
>library(mvtnorm)

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")
```

S[1] = 1, S[2]=1, S[12]=0



S[1] = 4, S[2]=3, S[12]=2



4.1 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 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

4.2 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)

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

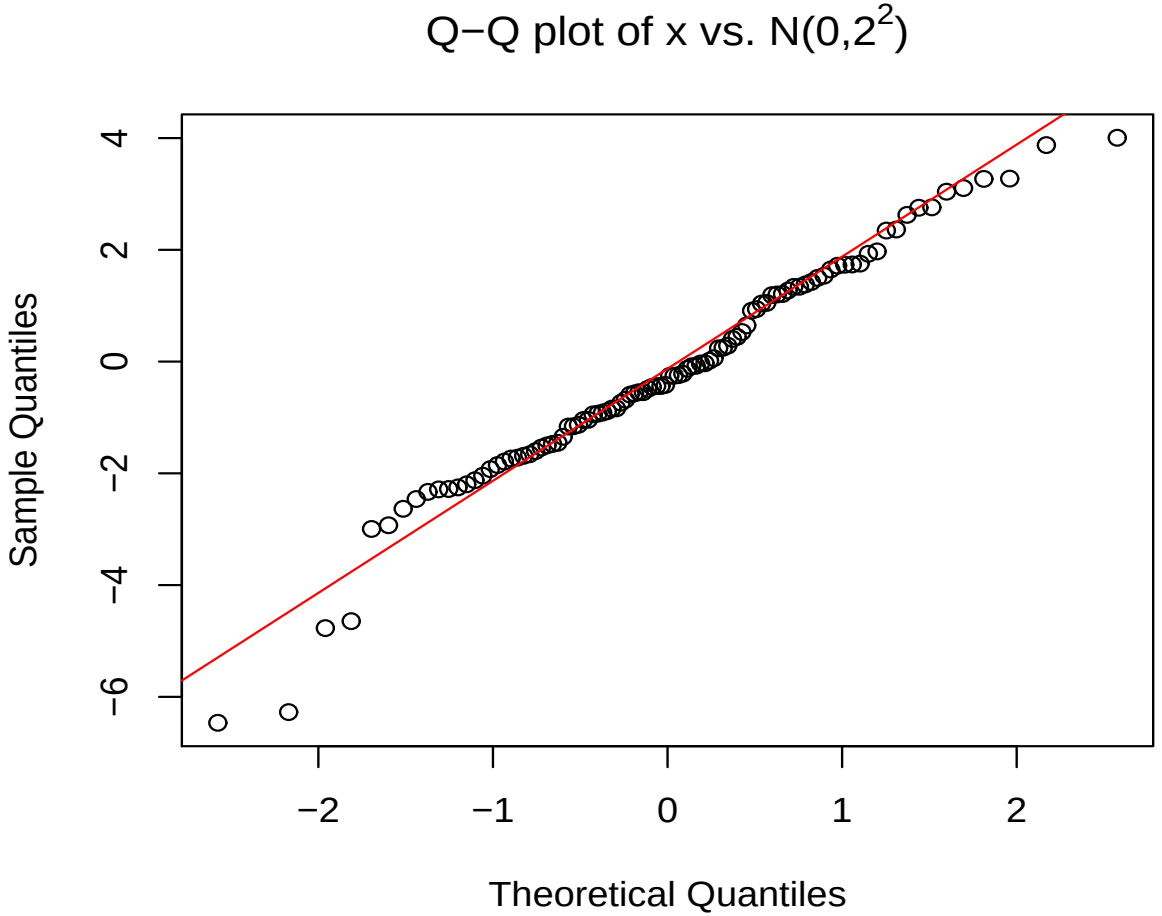
(b) Conditional distribution of $\mathbf{x}_1|_{\mathbf{x}_2} \sim N_q(\boldsymbol{\mu}_{1|2}, \Sigma_{11|2})$, where

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

4.3 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")
```


Q−Q plot of x vs. N(0,2^2)
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−6
−2
−1
0
1
2

- 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

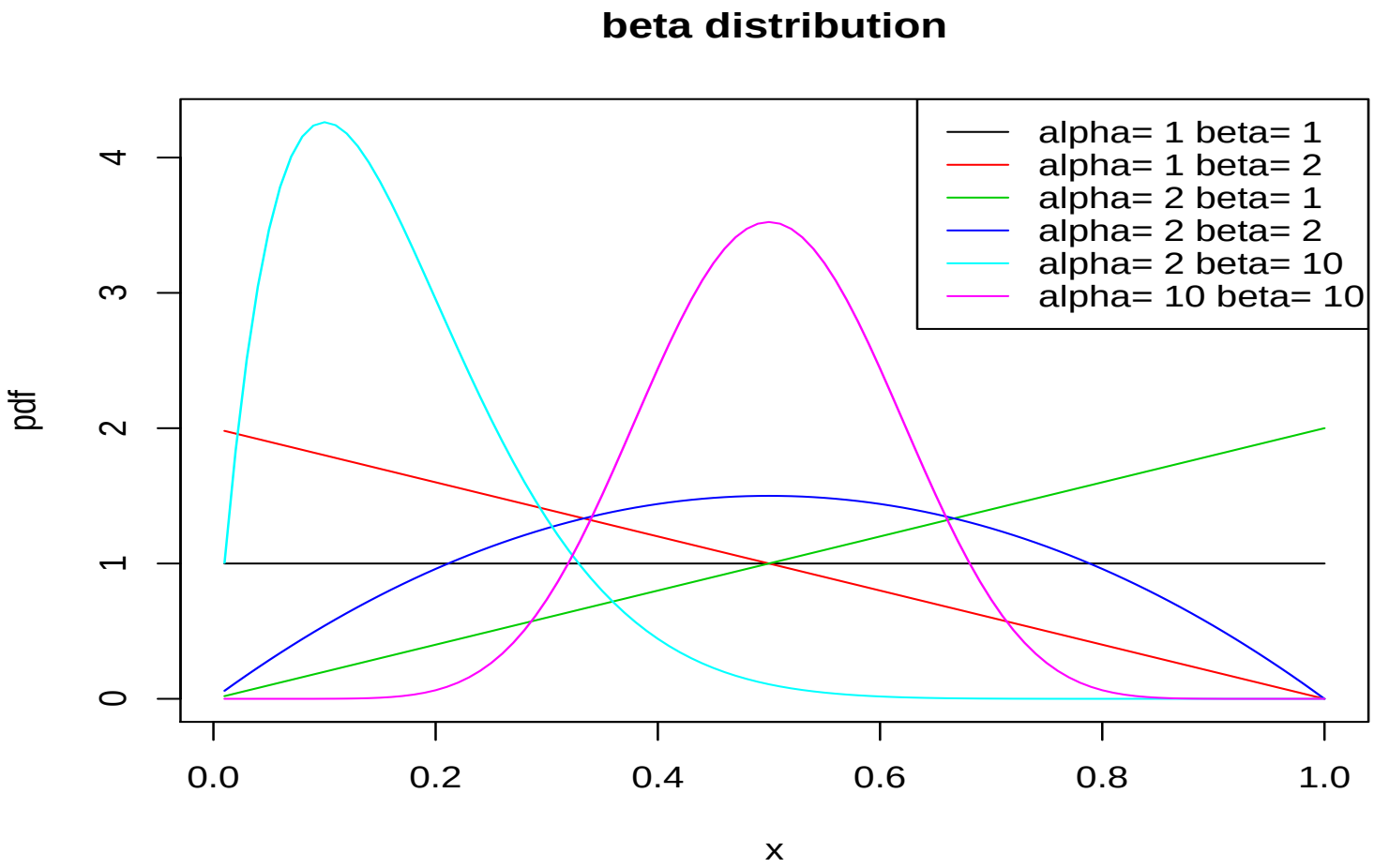
$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

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

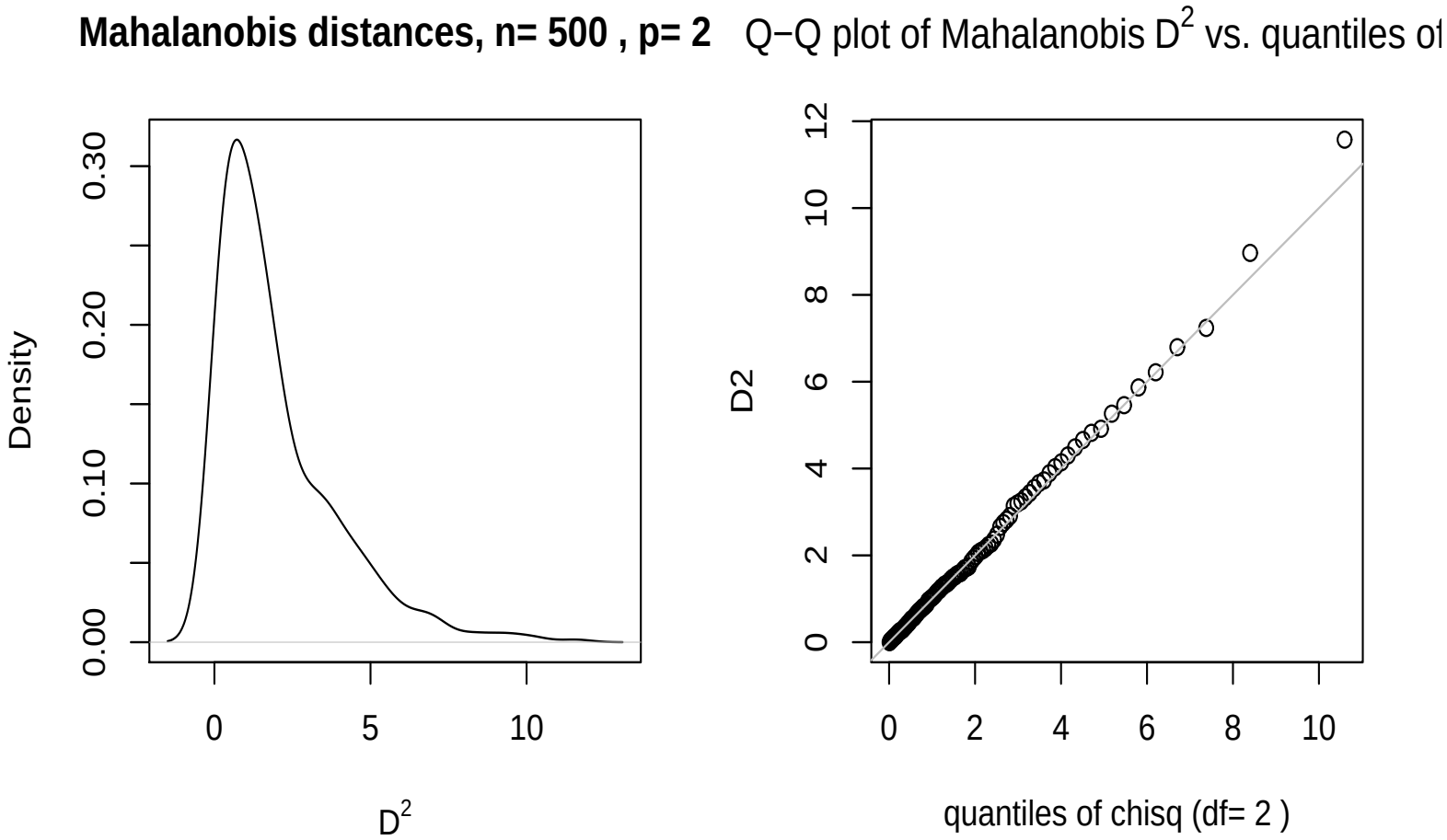
$$b_i = \frac{nD_i^2}{(n-1)^2} \sim Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$$

  (참고) Beta(α, β)분포의 pdf는

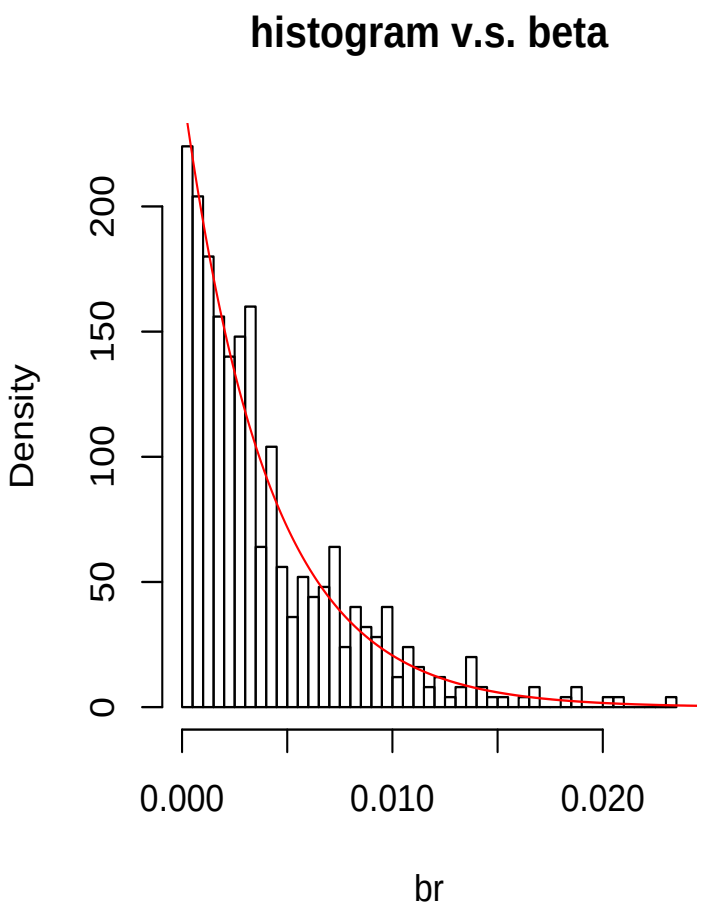
$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1.$$

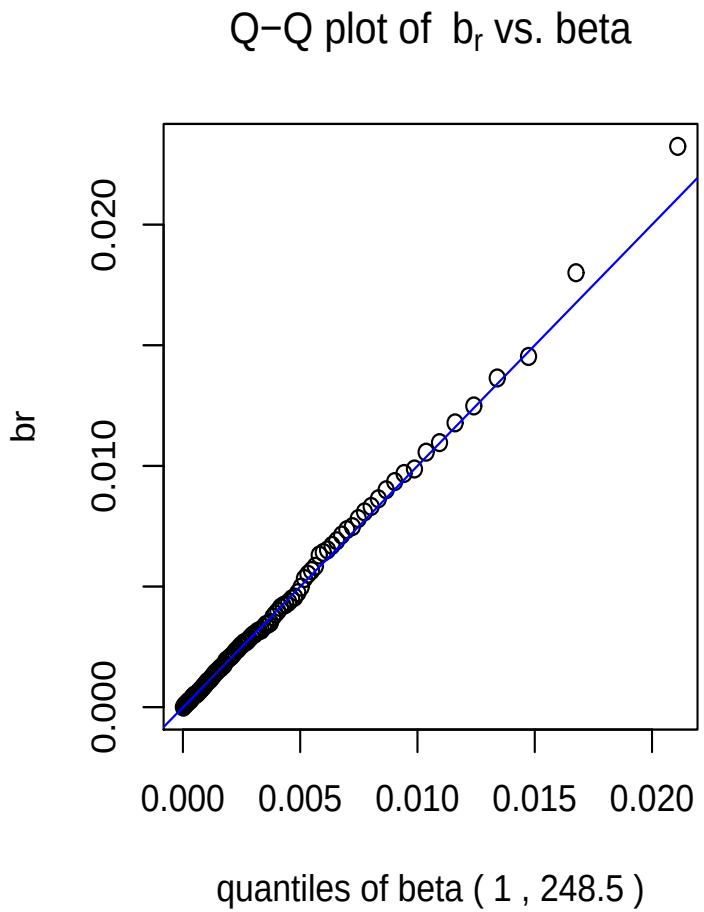

beta distribution
alpha= 1 beta= 1
alpha= 1 beta= 2
alpha= 2 beta= 1
alpha= 2 beta= 2
alpha= 2 beta= 10
alpha= 10 beta= 10
pdf
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
x

```
#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))
```


Mahalanobis distances, n= 500 , p= 2
Density
0.00
0.10
0.20
0.30
0
5
10
D2
Q−Q plot of Mahalanobis D2 vs. quantiles of
D2
0
2
4
6
8
10
12
0
2
4
6
8
10
quantiles of chisq (df= 2)

```
#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")
```


histogram v.s. beta
Density
0
50
100
150
200
0.000
0.010
0.020
br


Q−Q plot of b_r vs. beta
br
0.000
0.010
0.020
0.000
0.005
0.010
0.015
0.020
quantiles of beta (1 , 248.5)

- 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.

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

4.4 다변량자료의 변수변환 (Box-Cox transformation)