다변량정규분포

Jinseog Kim

October 8, 2007

1 다변량정규분포

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

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

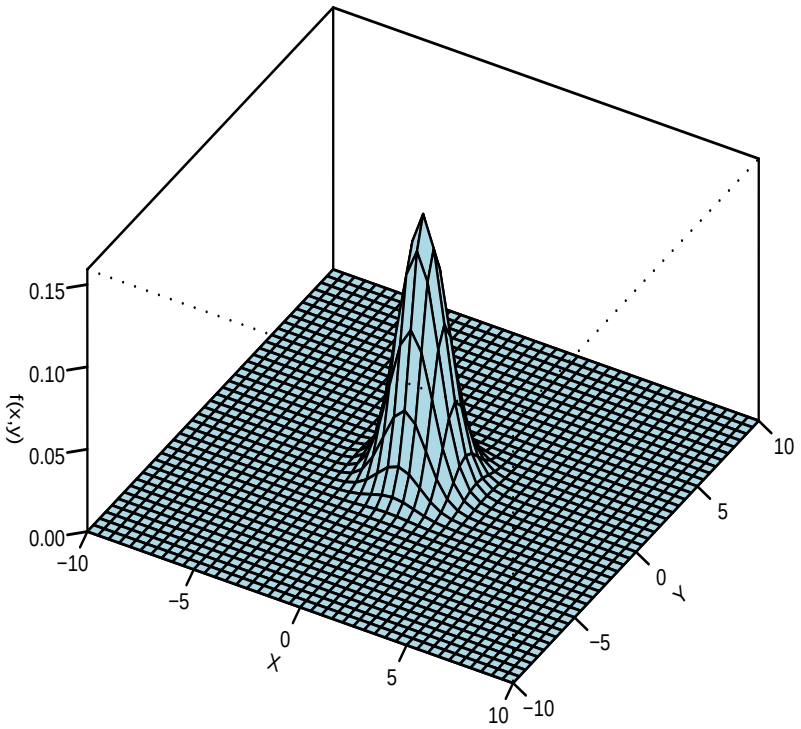
로 표현하고 이에 대한 확률밀도함수는

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

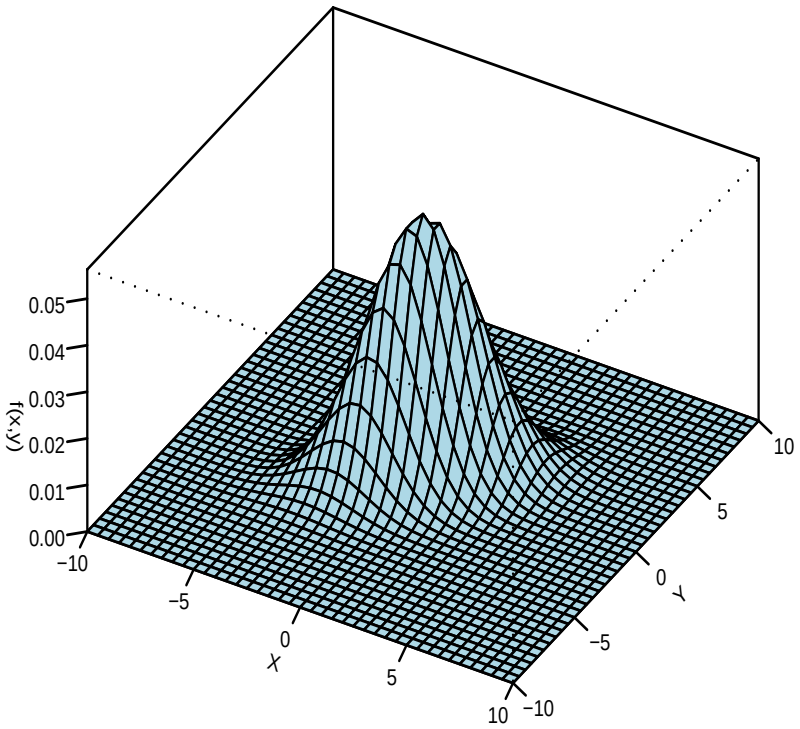
```
>library(mvtnorm)

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



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

1.2 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)

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

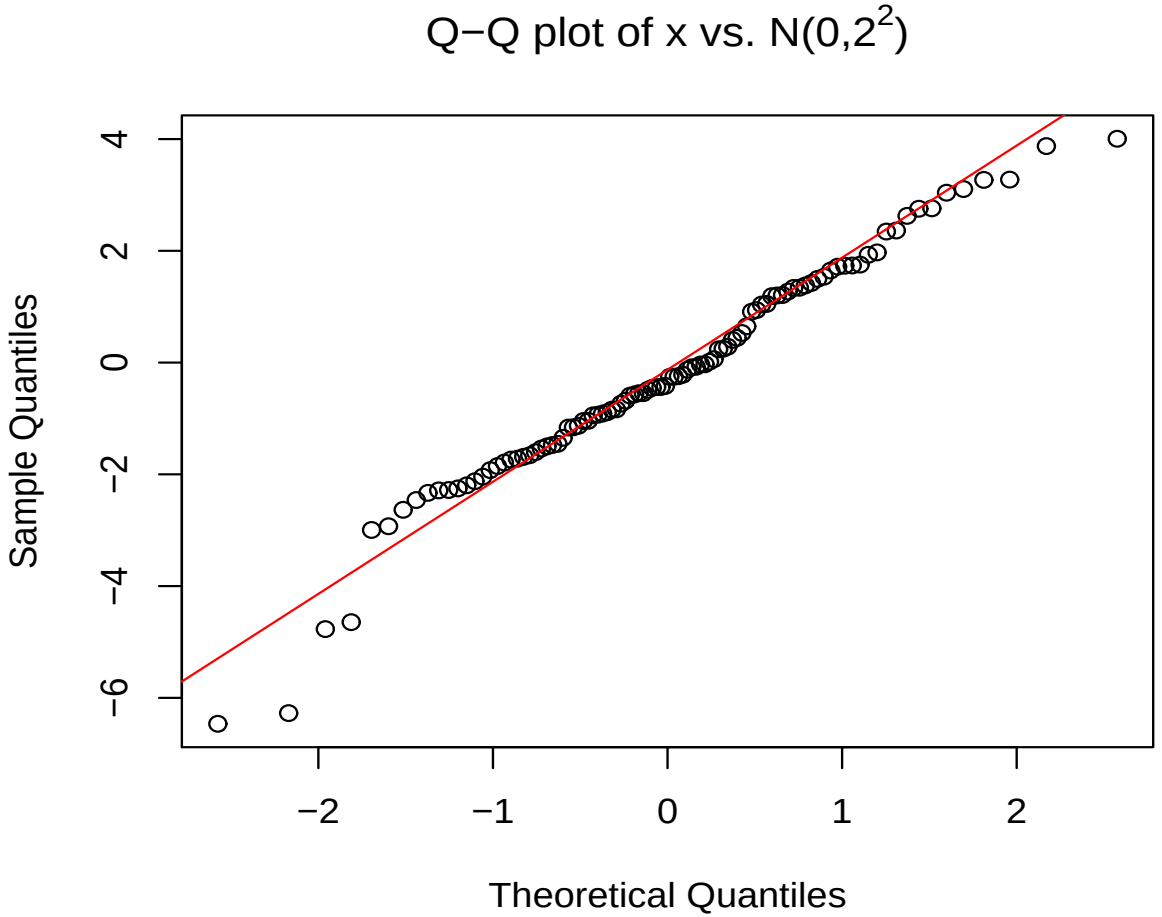
(b) Conditional distribution of $\mathbf{x}_1|_{\mathbf{x}_2} \sim N_q(\boldsymbol{\mu}_{1|2}, \Sigma_{11|2})$, where

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

1.3 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

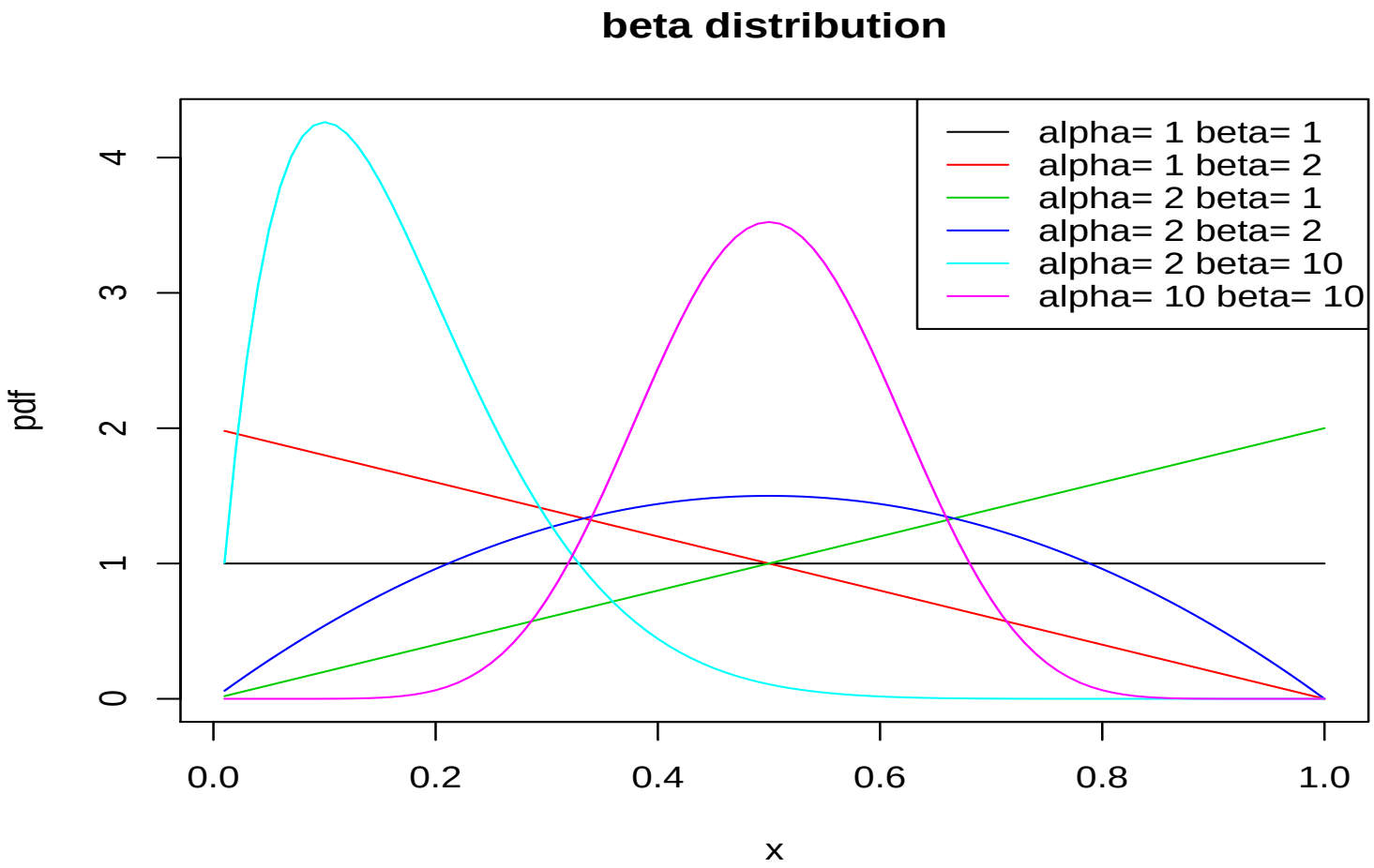
$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}).$$

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

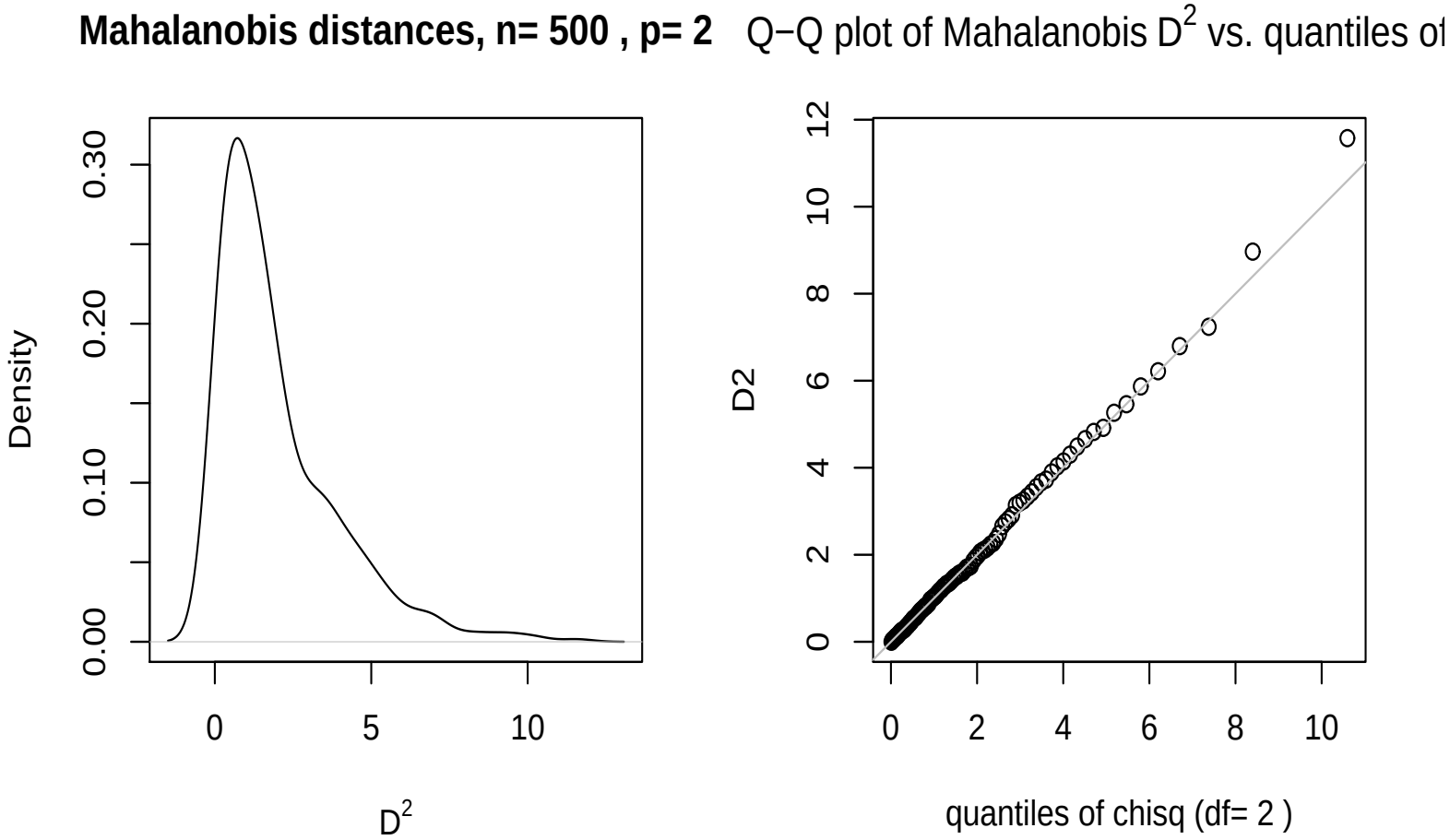
$$b_i = \frac{nD_i^2}{(n-1)^2} \sim Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$$

  (참고) Beta(α, β)분포의 pdf는

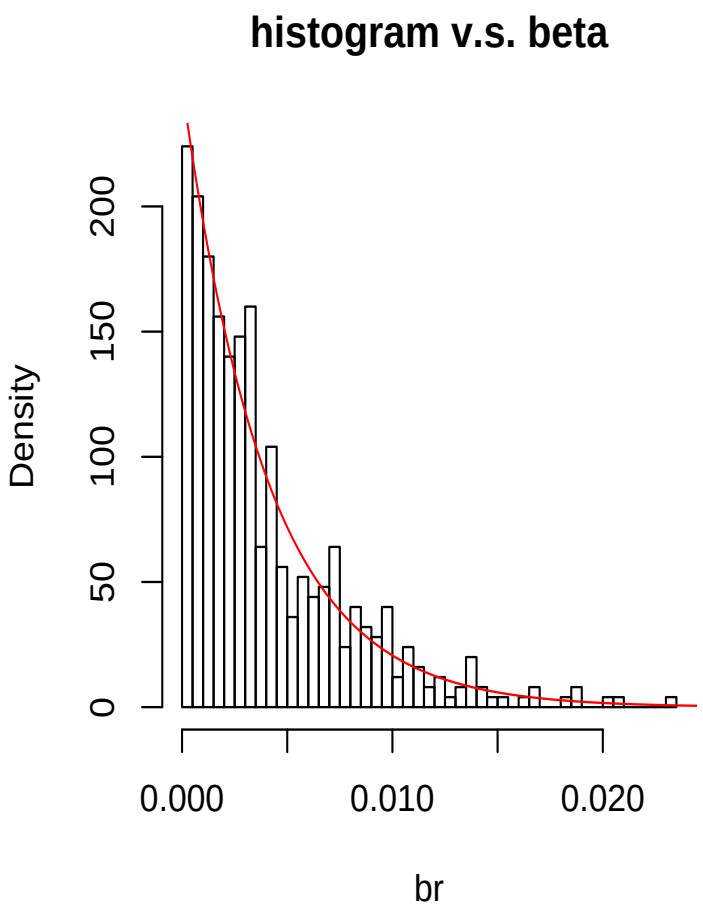
$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1.$$

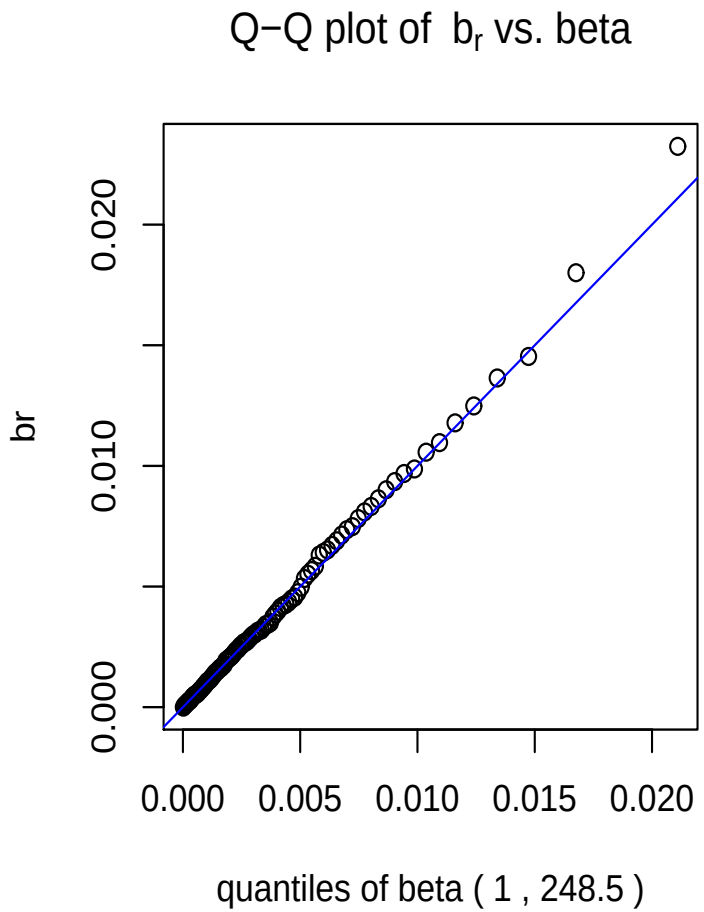

beta distribution
alpha= 1 beta= 1
alpha= 1 beta= 2
alpha= 2 beta= 1
alpha= 2 beta= 2
alpha= 2 beta= 10
alpha= 10 beta= 10
pdf
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
x

```
#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.

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

1.4 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

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

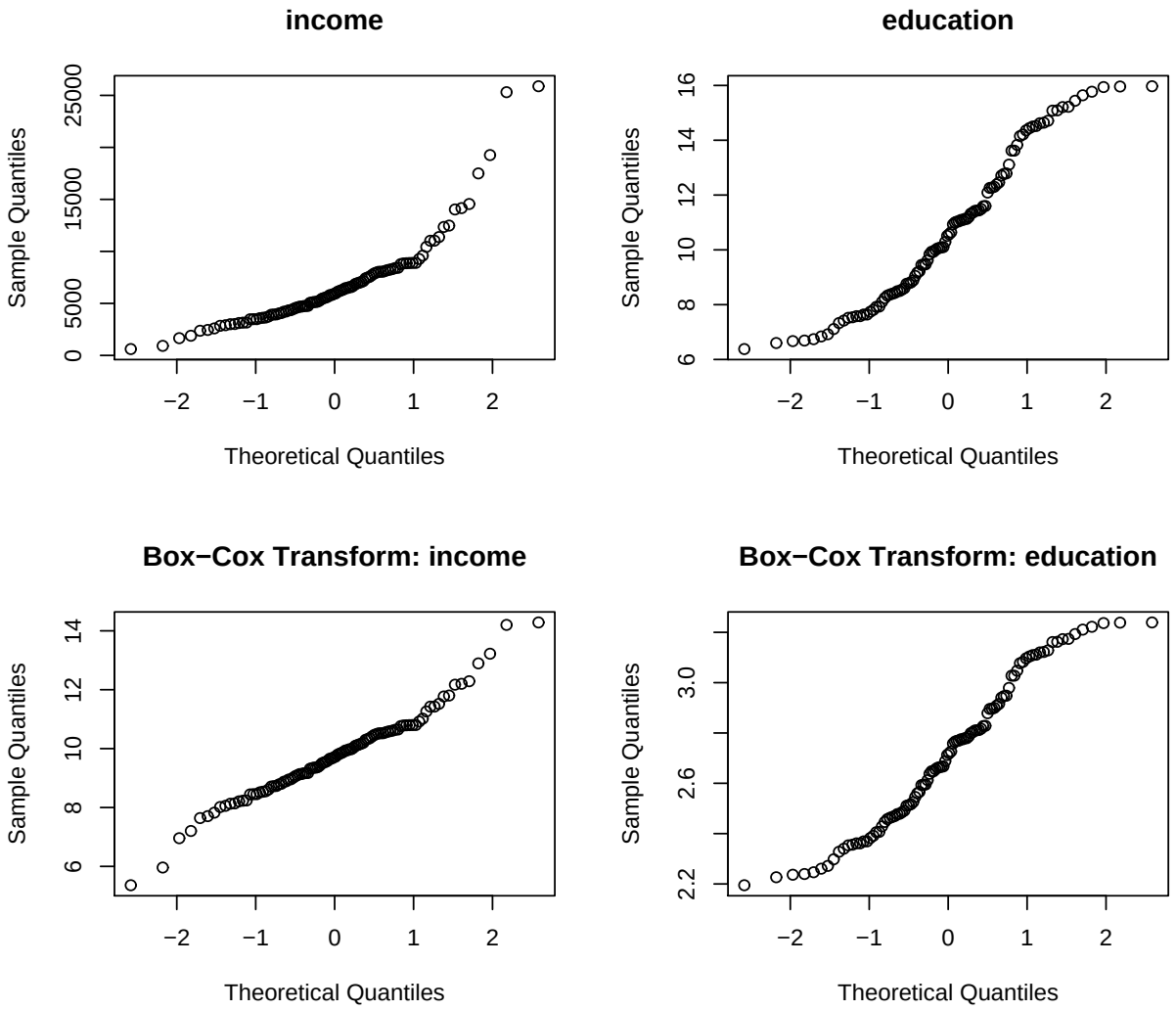
여기서 λ_j를 변환모수라고 부르며, 프로파일우도함수(Profile likelihood function)를 통해서 추정할 수 있다.

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")
```


income
Sample Quantiles
0
5000
15000
25000
−2
−1
0
1
2
Theoretical Quantiles
education
Sample Quantiles
6
8
10
12
14
16
−2
−1
0
1
2
Theoretical Quantiles
Box−Cox Transform: income
Sample Quantiles
6
8
10
12
14
−2
−1
0
1
2
Theoretical Quantiles
Box−Cox Transform: education
Sample Quantiles
2.2
2.6
3.0
−2
−1
0
1
2
Theoretical Quantiles