Олон хэмжээст өгөгдлийн статистик шинжилгээ

2024/10/17

Агуулга

Лекц I
Удиртгал

1 Хичээлийн агуулга

Хичээлийн агуулга
Үндсэн шинжилгээнүүд

1. Гол хэсгийн шинжилгээ (Principal Component Analysis)
2. Факторын шинжилгээ (Factor Analysis)
3. Кластерын шинжилгээ (Cluster Analysis)
4. Дискриминантын шинжилгээ (Discriminant Analysis)
5. Олон хэмжээст координатын шинжилгээ (Principal Coordinate Analysis эсвэл Multidimensional Scaling)
6. Конжойнт шинжилгээ (Conjoint Analysis)

7. Каноник корреляцын шинжилгээ (Canonical Correlation Analysis)
8. Хамтын тархалтын холбоо хамаарлын шинжилгээ (Correspondence Analysis)

Бусад

9. Олон хэмжээст хэвийн тархалт (Multivariate Normal Distribution)
10. Копула (Copula)

2 Үндсэн шинжилгээнүүд

Хэрэглээ

1. Үзэгдэл процесст нөлөөлөх хүчин зүйлсийг олж илрүүлэх
2. Объектуудыг ангилах, ялгаатай байдлыг олж тогтоох
3. Хувьсагчдыг ангилах
4. Хувьсагчдын холбоо хамаарлыг хэмжих, задлан шинжлэх
5. Хувьсагчдын холбоо хамаарлын зүй тогтлыг тогтоох
6. Санамсаргүй хувьсагчдын хамтын тархалтыг олж тогтоох

Гол хэсгийн шинжилгээ
Principal Component Analysis

Санамсаргүй хувьсагчдын холбоо хамаарал угтаа хэчнээн хүчин зүйлээр тодорхойлогдож байгааг олж тогтооход ашигладаг.

$$(X_1, \ldots, X_p)^T \longmapsto (Y_1, \ldots, Y_k)^T$$

$$k = ? : \ \ |DY| \approx |DX|$$

Энд

- $Y_1, \ldots, Y_k$ хувьсагчид хамааралгүй
- $DY_1 \geq \ldots \geq DY_k$
- $k \leq p$

Факторын шинжилгээ
Factor Analysis

Санамсаргүй хувьсагчдад яг k ширхэг хүчин зүйл буюу хувьсагч нөлөөлдөг гэж үзээд тэдгээр хүчин зүйлс болон холбогдох бусад зүйлсийг олно.

$$(X_1, \ldots, X_p)^T = L \cdot (f_1, \ldots, f_k)^T$$

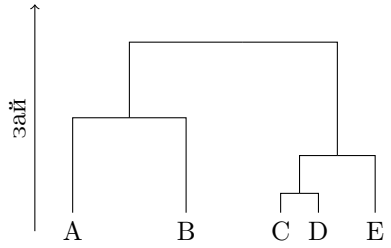
$$L = ?, \quad F = ?$$

Энд

- $F = (f_1, \ldots, f_k)^T$ факторууд хамааралгүй
- $Ef = 0$, $DF = I_k$
- L нь шугаман илэрхийллийн коэффициентуудаас тогтох матриц
- $k \leq p$

Кластерын шинжилгээ
Cluster Analysis

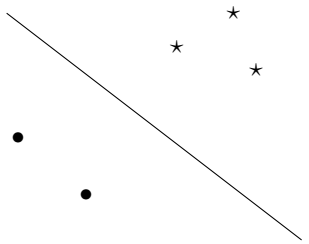
Хувьсагчид эсвэл объектуудын адил төст болон ялгаатай байдлыг олж илрүүлэх, хэчнээн ангид хувааж болохыг тогтоох эсвэл өгсөн тооны ангид хэрхэн хуваарилахыг тогтоох зэрэгт ашигладаг.



Зураг 1: Дендрограмм

Дискриминантын шинжилгээ
Discriminant Analysis

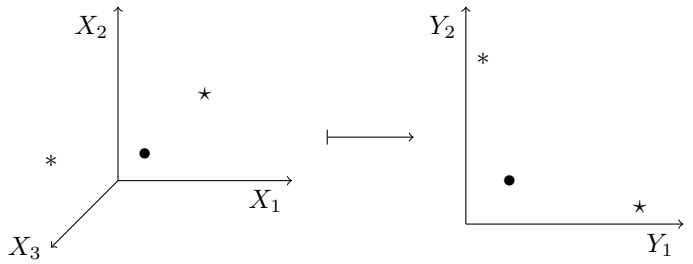
Бүлгүүдийн заагийг олж тогтоох зэрэгт ашигладаг.



Зураг 2: Дискриминантын шулуун

Олон хэмжээст координатын шинжилгээ
Principal Coordinate Analysis эсвэл Multidimensional Scaling

Их хэмжээст огторгуй дахь цэгүүдийг бага хэмжээст огторгуйд буулгах буюу проекцлох, зайн матрицаас цэгүүдийн координатыг сэргээн олох зэрэгт ашиглана.



Зураг 3: Хэмжээс багасгах

Конжойнт шинжилгээ
Conjoint Analysis

Регрессийн загвар дээр үндсэлсэн шинжилгээ бөгөөд нэн ялангуяа маркетингийн судалгаанд их хэрэглэдэг.

		X_2	
		1	2
X_1	1	2	1
	2	6	3
	3	5	4

(a) Өгөгдөл

		X_2	
		1	2
X_1	1	2.1	0.8
	2	5.1	3.8
	3	5.1	3.8

(b) Үр дүн

Хүснэгт 1: Шинжилгээний өгөгдөл ба эцсийн үр дүн

Каноник корреляцын шинжилгээ
Canonical Correlation Analysis

Хоёр санамсаргүй вектор хоорондын холбоо хамаарлыг хэмжихэд ашиглана.

$$\rho(X, Y)$$

энд

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} \qquad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad m, n \geq 2$$

Хамтын тархалтын холбоо хамаарлын шинжилгээ
Correspondence Analysis

Чанарын хувьсагчдын уялдаа холбоог задлан шинжлэхэд ашигладаг.

3 Шаардагдах суурь мэдлэг

Шаардагдах мэдлэг, чадвар

1. Матриц, түүн дээрх үйлдэл, чанар, хувийн утга болон хувийн вектор

		X		
		a_1	a_2	a_3
Y	b_1	1	3	2
	b_2	2	4	0

Хүснэгт 2: (X, Y) санамсаргүй векторын эх олонлогоос авсан түүврийн хамтын давтамжийн хүснэгт

2. Олон хэмжээст тархалт, тухайн тархалт, нөхцөлт тархалт
3. Момент, ковариацийн матриц
4. Магадлалын онол, математик статистик болон математикийн бусад суурь мэдлэг чадвар
5. R програм дээр ажиллах чадвар

Матрицын хувийн утга болон хувийн вектор

$A_{(p\times p)}$ нь квадрат матриц байг.

$$A\gamma = \lambda\gamma$$

байх λ скаляр тогтмол ба γ вектор оршин байвал эдгээрийг харгалзан A матрицын *хувийн утга* болон *хувийн вектор* гэнэ.

`A` матрицын хувийн утгууд болон хувийн векторуудыг дараах байдлаар олно.

```
eig <- eigen(A)
eig$values    # хувийн утгууд
eig$vectors   # хувийн векторууд
```

Олон хэмжээст тархалт

$X = (X_1, \ldots, X_p)^T$ нь санамсаргүй вектор байг.

$$F_X(x) = P(X < x) = P(X_1 < x_1, \ldots, X_p < x_p)$$

функцийг X санамсаргүй векторын *хамтын тархалтын функц* гэнэ. Харин

$$F_X(x) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_p} f(u_1, \ldots, u_p) du_1 \ldots du_p$$

$$\int_{\mathbb{R}^p} f_X(u) du = 1$$

нөхцөл хангах $f_X(x) \geq 0$: функцийг *хамтын нягтын функц* гэдэг.

Тухайн тархалт

$X = (X_1, X_2)^T$ санамсаргүй векторын хувьд $X_1 \in \mathbb{R}^k$ ба $X_2 \in \mathbb{R}^{p-k}$ байг. Тэгвэл X санамсаргүй векторын X_1 дэд векторын тухайн тархалтын функцийг

$$F_{X_1}(x_1) = P(X_1 < x_1) = \lim_{x_{k+1},\ldots,x_p \to +\infty} F_X(x_1, \ldots, x_k, x_{k+1}, \ldots, x_p)$$

байдлаар олно. Харин тухайн нягтын функцийг нь

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f(x_1, x_2) dx_{k+1} \ldots dx_p$$

интеграл бодож олно.

Нөхцөлт тархалт

X санамсаргүй вектор дахь зарим санамсаргүй хувьсагчдаас тогтох X_1 дэд векторын утгыг $X_1 = x_1$ гэж бэхэлсэн байг. Тэгвэл энэхүү нөхцөлд X_2 санамсаргүй векторын нөхцөлт нягтыг

$$f(x_2|x_1) = \frac{f_X(x_1, x_2)}{f_{X_1}(x_1)}$$

байдлаар олно. Тухайн тохиолдолд X_2 хувьсагч X_1 хувьсагчаас хамааралгүй үед

$$f_{X_2}(x_2|x_1) = f_{X_2}(x_2)$$

нөхцөл биелнэ.

Санамсаргүй векторын математик дундаж

Санамсаргүй векторын дундаж утга буюу математик дунджийг дараах байдлаар тодорхойлдог.

$$EX = \mu = \begin{pmatrix} EX_1 \\ \vdots \\ EX_p \end{pmatrix}$$

Чанар 1. 1. $a, b \in \mathbb{R}$ бол $E(aX + bY) = aEX + bEY$

2. $A_{(q\times p)}$ бодит тоон матриц бол $E(AX) = AEX$

3. X ба Y хамааралгүй бол $E(XY^T) = EXEY^T$

Санамсаргүй векторын ковариацийн матриц

Санамсаргүй векторын "дисперс" буюу ковариацийн матрицыг дараах байдлаар тодорхойлдог.

$$DX = \Sigma = \Sigma_{XX} = E(X - \mu)(X - \mu)^T = E\left((X_i - \mu_i)(X_j - \mu_j)\right)_{i,j=1,\ldots,p}$$
$$= \begin{pmatrix} D(X_1) & \operatorname{cov}(X_1, X_2) & \ldots & \operatorname{cov}(X_1, X_p) \\ \operatorname{cov}(X_2, X_1) & D(X_2) & \ldots & \operatorname{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p, X_1) & \operatorname{cov}(X_p, X_2) & \ldots & D(X_p) \end{pmatrix}$$

Чанар 2. $\Sigma^T = \Sigma$ буюу ковариацийн матриц тэгш хэмтэй.

Хоёр өөр санамсаргүй векторын ковариацийн матриц

Хоёр өөр санамсаргүй векторын ковариацийн матриц дараах хэлбэртэй байна.

$$\begin{aligned}\operatorname{cov}(X,Y) = \Sigma_{XY} &= E(X-\mu)(Y-\nu)^T = E\left((X_i-\mu_i)(Y_j-\nu_j)\right)_{i=1,\dots,p;\, j=1,\dots,q} \\ &= \begin{pmatrix} \operatorname{cov}(X_1,Y_1) & \operatorname{cov}(X_1,Y_2) & \dots & \operatorname{cov}(X_1,Y_q) \\ \operatorname{cov}(X_2,Y_1) & \operatorname{cov}(X_2,Y_2) & \dots & \operatorname{cov}(X_2,Y_q) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p,Y_1) & \operatorname{cov}(X_p,Y_2) & \dots & \operatorname{cov}(X_p,Y_q) \end{pmatrix}\end{aligned}$$

Чанар 3. 1. $\operatorname{cov}(X,Y) = E(XY^T) - EXEY^T$

2. X ба Y хамааралгүй бол $\operatorname{cov}(X,Y) = 0$

R програм дээрх олон хэмжээст өгөгдлийн хувьд түүврийн дундаж ба ковариацийн матриц олох

R програм дээрх матриц, датафрейм, тиббл зэрэг хэлбэртэй олон хэмжээст өгөгдлийн хувьд түүврийн дундаж утгын вектор болон түүврийн ковариацийн матрицыг дараах хоёр функцийн тусламжтай олно.

```
colMeans(X)
cov(X)
```

Энд `X` нь матриц, датафрейм, тиббл зэрэг хүснэгт хэлбэртэй өгөгдлийн төрөлд харгалзах хувьсагч юм.

Лекц II

Олон хэмжээст тархалт

1 Санамсаргүй векторын математик дундаж олох

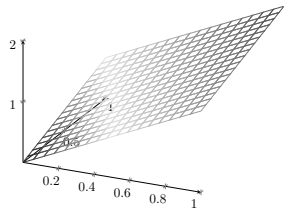
Жишээ болгон авч үзэх хоёр хэмжээст тархалт

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} \dfrac{4x_1+2x_2}{3}, & 0 \le x_1 \le 1,\ 0 \le x_2 \le 1 \\ 0, & \text{бусад} \end{cases} \qquad (*)$$

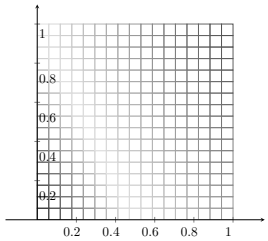
нягттай $X = (X_1, X_2)^T$ санамсаргүй вектор авч үзье.

$(*)$ нягттай санамсаргүй утгууд гарган авах

```
set.seed(0)
n <- 100000
X <- matrix(nrow = n, ncol = 2)
for (i in 1:n) {
  repeat {
    u <- runif(n = 1)
    Y <- runif(n = 2)
```



(a) нягтын функцийн график



(b) эгц дээрээс нь

Зураг 4: (∗) нягтын хэлбэр

```
    if (u < (4 * Y[1] + 2 * Y[2]) / 3 / 2) {
      X[i,] <- Y
      break
    }
  }
}
head(X)
plot(X, asp = 1, cex = 0.1, xlim = c(0,1), ylim = c(0,1), xlab = "X1",
    ylab = "X2")
```

Санамсаргүй векторын математик дундаж олох

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} f_{X_1,X_2}(x_1,x_2)dx_2 = \int_0^1 \frac{4x_1+2x_2}{3}dx_2 = \frac{4x_1+1}{3}$$
$$EX_1 = \int_{-\infty}^{+\infty} x_1 f_{X_1}(x_1)dx_1 = \int_0^1 x_1\frac{4x_1+1}{3}dx_1 = \frac{11}{18} \approx 0.611$$
$$f_{X_2}(x_2) = \int_{-\infty}^{+\infty} f_{X_1,X_2}(x_1,x_2)dx_1 = \int_0^1 \frac{4x_1+2x_2}{3}dx_1 = \frac{2+2x_2}{3}$$
$$EX_2 = \int_{-\infty}^{+\infty} x_2 f_{X_1}(x_1)dx_2 = \int_0^1 x_2\frac{2+2x_2}{3}dx_2 = \frac{5}{9} \approx 0.555$$

Ийнхүү $EX = \left(\frac{11}{18}, \frac{5}{9}\right)^T \approx (0.611, 0.555)^T$ үр дүнд хүрлээ.

2 Ковариацийн матриц

Ковариацийн матрицын чанар

Чанар 4. 1. a вектор бол $\operatorname{cov}(a^T X) = a^T \operatorname{cov}(X)a$

2. a вектор бол $\operatorname{cov}(X + a) = \operatorname{cov}(X)$

3. A матриц бол $\operatorname{cov}(AX) = A \operatorname{cov}(X)A^T$

4. $\operatorname{cov}(X + Y, Z) = \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z)$

5. $\operatorname{cov}(X+Y) = \operatorname{cov}(X) + \operatorname{cov}(X,Y) + \operatorname{cov}(Y,X) + \operatorname{cov}(Y)$
6. $\operatorname{cov}(AX, BY) = A\operatorname{cov}(X,Y)B^T$

Жишээ 1. A матриц бол $\operatorname{cov}(AX) = A\operatorname{cov}(X)A^T$ чанарыг батал.

Баталгаа

$$\begin{aligned}
\operatorname{cov}(AX) &= E(AX - E(AX))(AX - E(AX))^T \\
&= E(AX - AEX)(AX - AEX)^T \\
&= E(A(X - EX)((AX)^T - (AEX)^T)) \\
&= AE(X - EX)(X^TA^T - (EX)^TA^T) \\
&= AE(X - EX)(X^T - (EX)^T)A^T \\
&= AE(X - EX)(X - EX)^TA^T \\
&= A\operatorname{cov}(X)A^T
\end{aligned}$$

□

Жишээ 2. $\operatorname{cov}(X+Y, Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$ чанарыг батал.

Баталгаа

$$\begin{aligned}
&\operatorname{cov}(X+Y, Z) = \\
&\quad = E(X + Y - E(X+Y))(Z - EZ)^{T\,1} \\
&\quad = E((X - EX) + (Y - EY))(Z - EZ)^T \\
&\quad = E((X - EX)(Z - EZ)^T + (Y - EY)(Z - EZ)^T) \\
&\quad = E(X - EX)(Z - EZ)^T + E(Y - EY)(Z - EZ)^T \\
&\quad = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)
\end{aligned}$$

□

Санамсаргүй векторын ковариацийн матриц олох

Жишээ 3. $(*)$ тархалттай санамсаргүй векторын ковариацийн матрицыг ол.

$$\begin{aligned}
DX_1 = EX_1^2 - (EX_1)^2 &= \int_{-\infty}^{+\infty} x_1^2 f_{X_1}(x_1)dx_1 - (EX_1)^2 \\
&= \int_0^1 x_1^2 \frac{4x_1+1}{3}dx_1 - \frac{11^2}{18^2} = \frac{4}{9} - \frac{121}{324} = \frac{23}{324} \approx 0.071
\end{aligned}$$

Үүнтэй төстэй байдлаар $DX_2 = \frac{13}{162} \approx 0.080$ гэж олдоно.

$$\begin{aligned}
\operatorname{cov}(X_1, X_2) = E(X_1X_2) - EX_1EX_2 &= \int_{\mathbb{R}^2} x_1x_2 f(x_1,x_2)dx_1dx_2 - EX_1EX_2 \\
&= \int_0^1\int_0^1 x_1x_2\frac{4x_1+2x_2}{3}dx_1dx_2 - \frac{11}{18}\cdot\frac{5}{9} = \frac{1}{3} - \frac{55}{162} = -\frac{1}{162} \approx -0.006
\end{aligned}$$

Ийнхүү $\operatorname{cov}(X) = \begin{pmatrix} \frac{23}{324} & -\frac{1}{162} \\ -\frac{1}{162} & \frac{13}{162} \end{pmatrix} \approx \begin{pmatrix} 0.071 & -0.006 \\ -0.006 & 0.080 \end{pmatrix}$ болно.

Санамсаргүй векторын корреляцын матриц олох

Жишээ 4. $(*)$ тархалттай санамсаргүй векторын корреляцын матрицыг ол.

$\operatorname{cov}(X) = \begin{pmatrix} \frac{23}{324} & -\frac{1}{162} \\ -\frac{1}{162} & \frac{13}{162} \end{pmatrix}$ тул

$$\rho(X_1, X_2) = \frac{\operatorname{cov}(X_1, X_2)}{\sqrt{DX_1}\sqrt{DX_2}} = \frac{-\frac{1}{162}}{\sqrt{\frac{23}{324}}\sqrt{\frac{13}{162}}} = -\sqrt{\frac{2}{299}} \approx -0.082$$

бас $\rho(X_i, X_i) = 1$ байдаг тул

$$\rho(X) = \begin{pmatrix} 1 & -\sqrt{2/299} \\ -\sqrt{2/299} & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & -0.082 \\ -0.082 & 1 \end{pmatrix}$$

болно.

3 Нөхцөлт математик дундаж

Нөхцөлт математик дундаж

Тодорхойлолт 1.

$$E(X_1|X_2 = x_2) = \int x_1 f_{X_1|X_2}(x_1|x_2)dx_1$$

Чанар 5.
1. $E(E(X_1|X_2)) = EX_1$ (Бүтэн дунджийн томьёо)
2. $E(X_1 + X_2|X_3) = E(X_1|X_3) + E(X_2|X_3)$
3. X_1 ба X_2 хамааралгүй бол $E(X_1|X_2) = EX_1$
4. $E(\varphi(X_2)X_1|X_2) = \varphi(X_2)E(X_1|X_2)$

Жишээ 5. Бүтэн дунджийн томьёог батал.

Баталгаа

$$\begin{aligned} E(E(X_1|X_2)) &= \int E(X_1|X_2)f(x_2)dx_2 \\ &= \int \left(\int x_1 \cdot f(x_1|x_2)dx_1 \right) f(x_2)dx_2 \\ &= \int \left(\int x_1 \cdot f_{x_1,x_2}(x_1, x_2)dx_2 \right) dx_1 \\ &= \int x_1 \left(\int f(x_1, x_2)dx_2 \right) dx_1 \\ &= \int x_1 \cdot f(x_1)dx_1 \\ &= EX_1 \end{aligned}$$

□

Нөхцөлт математик дундаж олох

Жишээ 6. Хичээлийн эхэнд авч үзсэн хоёр хэмжээст санамсаргүй векторын хувьд $E(X_2|X_1)$ болон $E(X_1|X_2)$ нөхцөлт математик дунджуудийг нь ол.

$$\begin{aligned}
E(X_2|X_1 = x_1) &= \int x_2 f_{X_2|X_1}(x_2|x_1)dx_2 = \int x_2 \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}dx_2 \\
&= \int_0^1 x_2 \frac{\frac{4x_1+2x_2}{3}}{\frac{4x_1+1}{3}}dx_2 = \frac{2}{4x_1+1}\int_0^1 (2x_1x_2 + x_2^2)dx_2 \\
&= \frac{2}{3}\cdot\frac{3x_1+1}{4x_1+1} \\
E(X_1|X_2 = x_2) &= \frac{1}{6}\cdot\frac{4+3x_2}{1+x_2}
\end{aligned}$$

Нөхцөлт математик дундаж ба регрессийн шугаман загвар

Жишээний хувьд $E(X_2|X_1 = x_1) = \frac{2}{3}\cdot\frac{3x_1+1}{4x_1+1}$ шугаман бус хамаарал гарсан. Регрессийн шугаман загварын зүгээс хамаарлыг $E(X_2|X_1 = x_1) = a + bx_1$ байдалтай гэж үзнэ. Тэгвэл энэхүү шугаман бус хамаарал болон шугаман загварын хооронд хэр ялгаа гарах бол? Шугаман загварын a болон b параметрүүдийг симуляцийн аргаар өмнө гаргаж авсан хиймэл өгөгдөлд тулгуурлан олох буюу үнэлье. Үүнд `lm()` функц ашиглаж болно.

```
fit <- lm(formula = X[,2] ~ X[,1])
print(fit$coefficients)
```

Шугаман загвар ба жинхэнэ нөхцөлт математик дундаж хоёрын ялгааг диаграммаар дүрсэлж үзүүлнэ гэвэл R програм дээр дараах байдалтай код бичиж ажиллуулна.

```
plot.new(); dev.new(width = 5, height = 5, unit = "cm")
plot(X, asp = 1, cex = 0.2, xlim = c(0,1), ylim = c(0,1), xlab =
    "X1", ylab = "X2", col = "gray")
abline(reg = fit, col = "blue")
curve(expr = {2/3*(3*x+1)/(4*x+1)}, from = 0, to = 1, add = TRUE,
    col = "red")
```

4 Нөхцөлт корреляц ба тухайн корреляц

Нөхцөлт ковариац

Нөхцөлт ковариацийг дараах байдлаар тодорхойлдог.

$$\text{cov}(X_1|X_2 = x_2) = E((X_1 - E(X_1|X_2 = x_2))(X_1 - E(X_1|X_2 = x_2))^T|X_2 = x_2)$$

Энэ нь ердийн буюу нөхцөлт бус ковариацтэй төстэй боловч математик дундаж бодох бүртээ нөхцөл тооцдогоороо өөр юм. Мөн энд X_1 ба X_2 нь ерөнхий тохиолдолд санамсаргүй векторууд юм.

Чанар 6. 1. $\text{cov}(X_1) = E(\text{cov}(X_1|X_2)) + \text{cov}(E(X_1|X_2))$ /Бүтэн ковариацийн томьёо/

2. $\text{cov}(X_1|X_2) = E(X_1X_1^T|X_2) - E(X_1|X_2)E(X_1|X_2)^T$

Жишээ 7. Бүтэн ковариацийн томьёог батал.

Баталгаа

$$\begin{aligned}
&E(\text{cov}(X_1|X_2)) + \text{cov}(E(X_1|X_2)) = \\
&\quad = E\left\{E(X_1X_1^T|X_2) - E(X_1|X_2)E(X_1|X_2)^T{}^2\right\} + \\
&\quad\quad + \left[E\left\{E(X_1|X_2)E(X_1|X_2)^T\right\} - EX_1EX_1^T{}^3\right]{}^4 \\
&\quad = E(X_1X_1^T) - EX_1EX_1^T \\
&\quad = \text{cov}(X_1)
\end{aligned}$$

□

Нөхцөлт корреляц

$X_3 = x_3$ гэж бэхэлсэн үед X_1 ба X_2 санамсаргүй хувьсагчдын нөхцөлт ковариац ба нөхцөлт дисперсээр зохиох *нөхцөлт ковариацийн матриц*

$$\Sigma_{X_1,X_2|X_3=x_3} = \begin{pmatrix} D(X_1|X_3 = x_3) & \text{cov}(X_1, X_2|X_3 = x_3) \\ \text{cov}(X_1, X_2|X_3 = x_3) & D(X_2|X_3 = x_3) \end{pmatrix}$$

байна. Энд X_1, X_2 хоёр скаляр хувьсагч байх бол харин X_3 скаляр төдийгүй вектор байж болно. Улмаар

$$\rho(X_1, X_2|X_3 = x_3) = \frac{\text{cov}(X_1, X_2|X_3 = x_3)}{\sqrt{D(X_1|X_3 = x_3) \cdot D(X_2|X_3 = x_3)}}$$

байдлаар $X_3 = x_3$ гэж бэхэлсэн үеийн X_1 ба X_2 санамсаргүй хувьсагчдын хамаарлыг илэрхийлэх корреляцийн коэффициент зохиож болно. Ийм корреляцийг *нөхцөлт корреляц* гэнэ.

Тухайн корреляц

X_1 ба X_2 санамсаргүй хувьсагч тус бүрийг X_3 санамсаргүй вектороор илэрхийлсэн

$$X_1 = h(X_3) + U_1 \qquad X_2 = g(X_3) + U_2$$

загварууудын $U_1 = X_1 - h(X_3)$ ба $U_2 = X_2 - g(X_3)$ үлдэгдэл хоорондын

$$\text{cov}(X_1, X_2||X_3 = x_3) = \text{cov}(U_1, U_2) = \text{cov}(X_1 - h(X_3), X_2 - g(X_3))$$

ковариацийн коэффициентыг *тухайн ковариац* гэнэ. Өөрөөр хэлбэл энэ нь X_1 ба X_2 санамсаргүй хувьсагч тус бүр дээрх X_3 санамсаргүй вектор буюу хөндлөнгийн нийтлэг нөлөөг тооцож зайлуулсаны дараах үлдэгдэл хэсэг хоорондын ковариац юм. Мөн дээрх загварууудыг шугаманаар авдаг. Иймхүү тухайн ковариац олоход шугаман загвар ашигладаг.

Тухайн ковариацийн матриц улмаар *тухайн корреляцийн коэффициент* зэргийг дараах байдлаар тодорхойлно.

$$\Sigma_{X_1,X_2||X_3=x_3} = \begin{pmatrix} D(X_1||X_3) & \operatorname{cov}(X_1,X_2||X_3) \\ \operatorname{cov}(X_1,X_2||X_3) & D(X_2||X_3) \end{pmatrix}$$

$$\rho(X_1,X_2||X_3) = \frac{\operatorname{cov}(X_1,X_2||X_3)}{\sqrt{D(X_1||X_3)\cdot D(X_2||X_3)}}$$

Хэрэв X_1, X_2, X_3 бүгд хамтдаа олон хэмжээст хэвийн тархалттай эсвэл бусад эллиптик тархалттай бас олон хэмжээст гипергеометр, олон хэмжээст урвасан гипергеометр, мультномиал, Дирихлейн тархалттай бол тухайн корреляц нь нөхцөлт корреляцтай давхцах боловч ерөнхийдөө энэ хоёр корреляц ялгаатай юм.

Хагас тухайн корреляц

X_1 ба X_2 санамсаргүй хувьсагчдын аль нэгийг тухайлбал X_2 хувьсагчийг X_3 санамсаргүй вектороор илэрхийлсэн

$$X_2 = h(X_3) + U_2$$

загварын $U_2 = X_2 - h(X_3)$ үлдэгдэл ба нөгөө санамсаргүй хувьсагч X_1 хооронд ын

$$\operatorname{cov}(X_1,U_2) = \operatorname{cov}(X_1, X_2 - h(X_3))$$

ковариацийн коэффициентыг *хагас тухайн ковариац* гэнэ. Мөн үүний дагуу зохиох

$$\frac{\operatorname{cov}(X_1,U_2)}{\sqrt{D(X_1)\cdot D(U_2)}}$$

корреляцийг *хагас тухайн корреляц* гэнэ.

5 Санамсаргүй векторын хувиргалт

Санамсаргүй векторын хувиргалт

$u : \mathbb{R}^p \to \mathbb{R}^p$ функцээр X санамсаргүй векторыг хувиргах замаар Y санамсаргүй вектор зохиож байгаа гэе.

$$Y = u(X)$$

Томьёо 1.

$$f_Y(y) = \operatorname{abs}(|J|) f_X(u^{-1}(y))$$

Энд $J = \left(\frac{\partial x_i(y)}{\partial y_j}\right)_{i,j=1,\ldots,p}$ бол Якобиан буюу Якобын матриц, $|J|$ бол J матрицын тодорхойлогч, $u^{-1}(y)$ бол $u(x)$ функцийн урвуу функц юм.

Харин A матриц, b векторын тусламжтай $Y = AX + b$ шугаман хувиргалтын хувьд дээрх томьёо дараах хэлбэртэй болно.

$$f_Y(y) = \operatorname{abs}(|A|^{-1}) f_X(A^{-1}(y-b))$$

Жишээ 8. X нь $(*)$ тархалттай бол $Y = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}$ санамсаргүй векторын нягтыг ол.

$Y = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = AX + b$ шугаман хувиргалт өгчээ. Иймд $f_Y(y) = \text{abs}(|A|^{-1}) f_X(A^{-1}(y-b))$ томьёо ашиглана. Энд $|A| = -2$, $\text{abs}(|A|^{-1}) = \frac{1}{2}$, $A^{-1} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$ тул

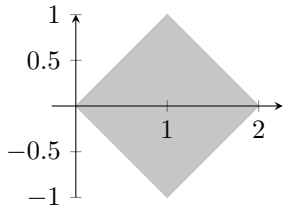
$$f_Y(y) = \frac{1}{2} f_X \left\{ \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\} = \frac{1}{2} f_X \left\{ \frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2} \right\}$$
$$= \frac{1}{2} \cdot \frac{4 \cdot \frac{y_1 + y_2}{2} + 2 \cdot \frac{y_1 - y_2}{2}}{3} = \frac{3y_1 + y_2}{6}$$

болно. Дээрх илэрхийлэл $y = (y_1, y_2)$ аргументын ямар утганд харгалзахыг дараагийн слайд дээр авч үзнэ.

X санамсаргүй векторын хувьд $0 \le x_1 \le 1$ ба $0 \le x_2 \le 1$ байсан тул $0 \le \frac{y_1+y_2}{2} \le 1$ ба $0 \le \frac{y_1-y_2}{2} \le 1$ буюу

$$|y_1 + y_2 - 1| \le 1 \text{ ба } |y_1 - y_2 - 1| \le 1$$

болно.



Зураг 5: Y санамсаргүй векторын авах утгууд

```
A <- matrix("data" = c(1,1,1,-1), "nrow" = 2); b <- c(0,0)
Y <- X
for (i in 1:nrow(Y))
  Y[i,] <- A %*% Y[i,] + b
plot(Y, asp = 1, cex = 0.2, xlim = c(0,2), ylim = c(-1,1), xlab =
    "Y1", ylab = "Y2", col = "gray")
```

Лекц III

Олон хэмжээст хэвийн тархалт I

1 Тодорхойлолт

Тодорхойлолт

Сэргээн санах нь 1 (Нэг хэмжээст хэвийн тархалт).

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$f_X(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

нягттай тархалтыг *олон хэмжээст хэвийн тархалт* гэнэ. Энд $|\Sigma|$ нь Σ матрицын тодорхойлогч юм. $X = (X_1, \ldots, X_p)^T$ санамсаргүй векторыг μ болон Σ параметрүүд бүхий олон хэмжээст хэвийн тархалттай гэхийг $X \sim N_p(\mu, \Sigma)$ гэж тэмдэглэнэ.

Олон хэмжээст хэвийн тархалтын параметрууд

$X = (X_1, \ldots, X_p)^T \sim N_p(\mu, \Sigma)$ байг. Тэгвэл тус тархалтын μ болон Σ параметрууд нь X санамсаргүй векторын дундаж утгын вектор болон ковариацийн матриц өөрөөр хэлбэл

$$\mu = EX = \begin{pmatrix}\mu_1 \\ \vdots \\ \mu_p\end{pmatrix} = \begin{pmatrix}EX_1 \\ \vdots \\ EX_p\end{pmatrix}$$

$$\Sigma = \operatorname{cov}(X) = E(X-\mu)(X-\mu)^T = \big(E(X_i-\mu_i)(X_j-\mu_j)\big)_{i,j=1,\ldots,p}$$

$$= \begin{pmatrix} \operatorname{cov}(X_1,X_1) & \operatorname{cov}(X_1,X_2) & \ldots & \operatorname{cov}(X_1,X_p) \\ \operatorname{cov}(X_2,X_1) & \operatorname{cov}(X_2,X_2) & \ldots & \operatorname{cov}(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p,X_1) & \operatorname{cov}(X_p,X_2) & \ldots & \operatorname{cov}(X_p,X_p) \end{pmatrix}$$

байна.

Олон хэмжээст хэвийн тархалтын нягтын функц болон тархалтын функцийн утга олох бэлэн функц

Нягтын функц

```
mvtnorm::dmvnorm(x, mean, sigma)
```

x хувьсагчийн утга, вектор эсвэл матриц хэлбэртэй байна.

mean дундаж утгын вектор

sigma ковариацийн матриц

Тархалтын функц

```
mvtnorm::pmvnorm(lower = -Inf, upper, mean, sigma)
```

lower доод хязгаар

upper дээд хязгаар

mean дундаж утгын вектор

sigma ковариацийн матриц

Жишээ 9. $\mu = (5, -4)^T$ дундаж утгын вектор болон $\Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$ ковариацийн матриц бүхий хоёр хэмжээст хэвийн тархалт авч үзье.

```
mu <- c(5,-4)
Sigma <- matrix(c(4,-1,-1,1), ncol = 2)
```

$f_{X_1,X_2}(3, -2)$

```
mvtnorm::dmvnorm(x = c(3,-2), mean = mu, sigma = Sigma)
```

$P(-\infty < X_1 < 3, -\infty < X_2 < -2)$

```
mvtnorm::pmvnorm(lower = -Inf, upper = c(3,-2), mean = mu, sigma
    = Sigma)
```

Хамааралгүй хувьсагчдын тархалт

$X_1, \ldots, X_p$ хувьсагчид хамааралгүй бол

$$\Sigma = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & 0 & \ldots & 0 \\ 0 & \operatorname{cov}(X_2, X_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \operatorname{cov}(X_p, X_p) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_p^2 \end{pmatrix}$$

байх ба улмаар

$$f_X(x) = \prod_{i=1}^{p} f_{X_i}(x_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{1}{2\sigma_i^2}(x_i - \mu_i)^2 \right\}$$

болно.

Хоёр хэмжээст хэвийн тархалт

ρ корреляцтай X_1 ба X_2 санамсаргүй хувьсагчдаас тогтох $X = (X_1, X_2)^T$ санамсаргүй векторын хоёр хэмжээст хэвийн тархалтын тэмдэглэгээг дэлгэрэнгүй бичвэл

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

хэлбэртэй байна. Харин хамтын нягтын функц нь

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \cdot \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

болно.

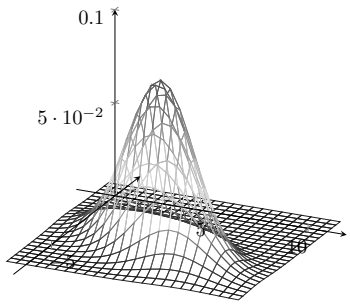
Жишээ 10. $\mu = (5, -4)$ дундаж утгын вектор болон $\Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$ ковариацийн матриц бүхий хоёр хэмжээст хэвийн тархалтын хамтын нягтын илэрхийллийг бичиж, графикийг нь зур.

$\mu = (\mu_1, \mu_2)^T = (5, -4)^T$ ба $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$ гэдгээс $\sigma_1 = 2$, $\sigma_2 = 1$, $\rho = -1/2$ болно. Иймд нягт нь

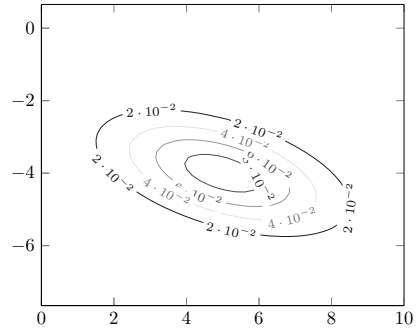
$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\sqrt{3}\pi} \exp\left\{ -\frac{1}{6}\left[(x_1 - 5)^2 + 2(x_1 - 5)(x_2 + 4) + 4(x_2 + 4)^2 \right] \right\}$$

болно. Харин графикийг нь дараагийн слайд дээр байгуулж үзүүлнэ.

Жишээ болгон авсан хоёр хэмжээст хэвийн тархалтын нягтын функцийн график, түүний түвшний шугамыг дараах зургаар харууллаа.



нягтын функцийн график



нягтын функцийн түвшний шугам

2 Хувиргалт ба симуляц

Махаланобисын хувиргалт

$X \sim N_p(\mu, \Sigma)$ санамсаргүй векторыг

$$Z = \Sigma^{-1/2}(X - \mu)$$

байдлаар хувиргавал олон хэмжээст стандарт хэвийн тархалттай өөрөөр хэлбэл

$$Z \sim N_p(0, I_p)$$

болно. Энд $Z = (Z_1, \ldots, Z_p)^T$, 0 нь тэг вектор, I_p нь p хэмжээст нэгж матриц юм. Иймд $Z_1, \ldots, Z_p$ хувьсагчид хамааралгүй бөгөөд нэг хэмжээст $N(0, 1)$ стандарт хэвийн тархалттай юм. Дээрх хувиргалтыг *Махаланобисын хувиргалт* гэдэг.

Олон хэмжээст хэвийн тархалттай санамсаргүй утга хиймлээр үүсгэх буюу симуляцлах

$Z \sim N_p(0, I_p)$ буюу $Z = (Z_1, \ldots, Z_p)^T$ бөгөөд $Z_1, \ldots, Z_p$ хувьсагчид хамааралгүй бөгөөд нэг хэмжээст $N(0,1)$ стандарт хэвийн тархалттай бол

$$X = \Sigma^{1/2} Z + \mu \sim N_p(\mu, \Sigma)$$

болно.

$$\begin{aligned}
&E(\Sigma^{1/2} Z + \mu) = E(\Sigma^{1/2} Z) + E\mu = \Sigma^{1/2} EZ + \mu = \mu \\
&\operatorname{cov}(\Sigma^{1/2} Z + \mu) \\
&\quad = E(\Sigma^{1/2} Z + \mu - E(\Sigma^{1/2} Z + \mu))(\Sigma^{1/2} Z + \mu - E(\Sigma^{1/2} Z + \mu))^T \\
&\quad = E(\Sigma^{1/2} Z)(\Sigma^{1/2} Z)^T = \Sigma^{1/2} E(ZZ^T)\Sigma^{1/2} {}^{5} = \Sigma^{1/2} I_p \Sigma^{1/2} = \Sigma
\end{aligned}$$

Олон хэмжээст хэвийн тархалттай санамсаргүй утга үүсгэх бэлэн функц

```
mvtnorm::rmvnorm(n, mean, sigma)
```

n үүсгэх санамсаргүй утгын тоо

mean дундаж утгын вектор

sigma ковариацийн матриц

```
MASS::mvrnorm(n, mu, Sigma)
```

n үүсгэх санамсаргүй утгын тоо

mu дундаж утгын вектор

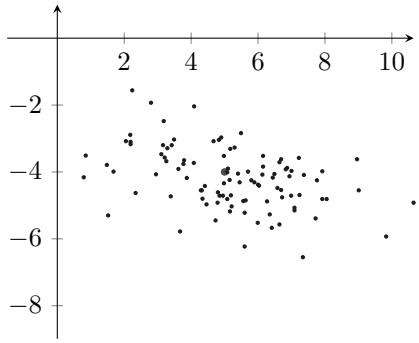
Sigma ковариацийн матриц

Жишээ 11.
```
MASS::mvrnorm(
 n = 100,
 mu = c(5,-4), Sigma = matrix(c(4,-1,-1,1), ncol = 2)
)
```

Хувиргалт

Теорем 1. $X \sim N_p(\mu, \Sigma)$ ба A нь p эрэмбийн квадрат, үл бөхөх матриц; b нь p хэмжээст вектор бол $Y = AX + b$ санамсаргүй вектор нь $N_p(A\mu + b, A\Sigma A^T)$ тархалттай байна.

Баталгаа
- $f_Y(y) = \operatorname{abs}(|A|^{-1}) f_X(A^{-1}(y - b))$
- $f_X(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$



- $\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right]_{x=A^{-1}(y-b)} = (A^{-1}(y-b)-\mu)^T\Sigma^{-1}(A^{-1}(y-b)-\mu) = (A^{-1}(y-b)-A^{-1}A\mu)^T\Sigma^{-1}(A^{-1}(y-b)-A^{-1}A\mu) = (y-(A\mu+b))^T(A^{-1})^T$[6]$\Sigma^{-1}A^{-1}(y-(A\mu+b)) = (y-(A\mu+b))^T(A^T)^{-1}\Sigma^{-1}A^{-1}$[7]$(y-(A\mu+b)) = (y-(A\mu+b))^T(A\Sigma A^T)^{-1}(y-(A\mu+b))$

□

3 Геометр агуулга

Геометр агуулга

Энэ хэсэгт $(x-\mu)^T\Sigma^{-1}(x-\mu) = d^2$ эллипсоидийн шинж чанарыг үзнэ.

Чанар 7. 1.

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = d^2$$

тэгшитгэлээр тодорхойлогдох эллипсоидод харгалзах цэгүүдийн хувьд $N_p(\mu,\Sigma)$ тархалтын нягт тогтмол байна.

2. $X \sim N_p(\mu,\Sigma)$ бол

$$U = (X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi_p^2$$[8]

байна.

Нэг дүгээр чанар илэрхий тул хоёр дугаар чанарыг л баталъя.

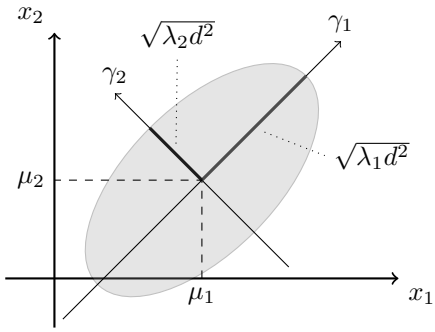
Баталгаа $X \sim N_p(\mu,\Sigma)$

$$\begin{aligned} U &= (X-\mu)^T\Sigma^{-1}(X-\mu) \\ &= (X-\mu)^T\Sigma^{-1/2}\Sigma^{-1/2}(X-\mu) \\ &= (\Sigma^{-1/2}(X-\mu))^T\Sigma^{-1/2}(X-\mu) \\ &= Y^TY \\ &= \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2 \end{aligned}$$[9][10]

□

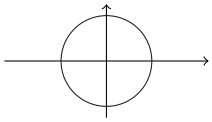
[6] $(A^T)^{-1} = (A^{-1})^T$
[7] $(AB)^{-1} = B^{-1}A^{-1}$
[8] p чөлөөний зэрэгтэй хи-квадрат тархалт

$(x-\mu)^T\Sigma^{-1}(x-\mu) = d^2$ эллипсоидийн их болон бага тэнхлэгүүдийн чиглэл болон хэмжээ

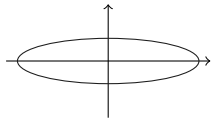


Энд λ_i болон γ_i нь Σ ковариацийн матрицын хувийн утга болон хувийн вектор юм. Мөн эллипсоидын "талбай" $S = \dfrac{2\pi^{p/2}}{p\Gamma(p/2)} d^p |\Sigma|^{1/2}$ байна.

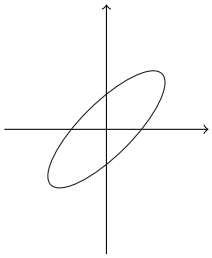
$(x-\mu)^T\Sigma^{-1}(x-\mu) = d^2$ эллипсоидийн хэлбэр болон их тэнхлэгийн чиглэл ба Σ матриц



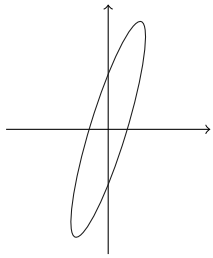
$DX_1 = DX_2,\ \rho(X_1, X_2) = 0$



$DX_1 = 4 \cdot DX_2,\ \rho(X_1, X_2) = 0$



$DX_1 = DX_2,\ \rho(X_1, X_2) = 0.5$



$DX_1 = 4 \cdot DX_2,\ \rho(X_1, X_2) = 0.5$

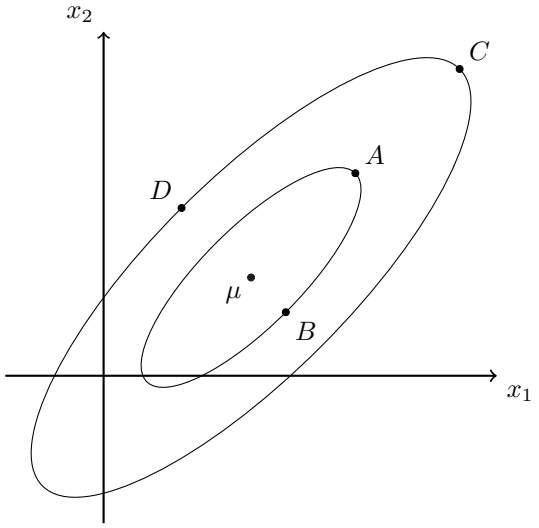
Махаланобисын зай

x цэгээс Σ ковариацийн матрицтай тархалтын төв μ хүртэлх

$$d(x, \mu) = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

зайг *Махаланобисын зай* гэдэг. $d(x, \mu) = 0$ байх гарцаагүй бөгөөд хүрэлцээтэй нөхцөл нь $x = \mu$ байх явдал бөгөөд харин x цэг эллипсоидын их тэнхлэгийн дагуу гадагш чиглэн хөдөлбөл тархалтын төвөөс хамгийн хурдан холдоно. Тархалтын төвөөс Махаланобисын утгаар ижил зайд байх цэгүүдийн олонлог

эллипсоид үүсгэх бөгөөд олон хэмжээст хэвийн тархалтын хувьд тэрхүү эллипсоид дээрх цэгүүд нь ижил нягттай буюу тус санамсаргүй векторын ижил боломжтой утгууд юм.



$$d(A,\mu) = d(B,\mu)$$
$$d(A,\mu) < d(D,\mu)$$
$$d(C,\mu) = d(D,\mu)$$

Махаланобисын зай олох

Махаланобисын зайг R програмын `stats` багц дахь `mahalanobis()` функцийн тусламжтай олж болно. Тус функц нь Махаланобисын зайн квадрат утгыг буцаадаг.

```
mahalanobis(x, center, cov)
```

x өгөгдөл бүхий матриц эсвэл датафрейм

center тархалтын төвийг заасан вектор

cov ковариацийн матриц

Махаланобисын зай ба олон хэмжээст хэвийн тархалттай өгөгдөл дэх онцгой утга илрүүлэх нь

$X \sim N_p(\mu, \Sigma)$ үед $(X-\mu)^T\Sigma^{-1}(X-\mu)$ хувьсагч p чөлөөний зэрэгтэй хи-квадрат тархалттай байдаг талаар энэ хэсгийн эхэнд үзсэн. Иймд тус хувьсагчийн хувьд $d^2(X,\mu) \sim \chi^2_p$ буюу Махаланобисын зайн квадратаар зохиогдох хувьсагч мөн адил хи-квадрат тархалттай юм. Үүнд үндэслэн олон хэмжээст хэвийн тархалттай хэмээн тооцож буй өгөгдөл дэх магадлал багатай буюу онцгой утгуудыг илрүүлдэг дараах томьёолол бүхий шинжүүр зохиож болно.

$$d^2(x,\mu) \geq \chi^2_{p,1-\alpha}$$

Энд x нь түүврийн элемент харин $\chi^2_{p,1-\alpha}$ нь p чөлөөний зэрэг бүхий хи-квадрат тархалтын $1-\alpha$ эрэмбийн квантил юм.

Онцгой утга илрүүлэхэд R програмын `rstatix` багц дахь `mahalanobis_distance()` функц ашиглаж болно. Тус функцийн хувьд шинжүүрийн итгэх түвшинг $\alpha = 0.001$ гэж сонгосон байдаг.

Жишээ 12. R програмын `datasets` багцад бэлэн байдаг, цахилдаг цэцгийн цоморлиг болон дэлбээний урт ба өргөний хэмжээг илэрхийлсэн өгөгдөл бүхий `iris3` массиваас Setosa зүйлд холбогдох мэдээллийг ялган авч улмаар онцгой утга буй эсэхийг тогтоо.

```
# Мэдээлэл ялгаж авах
X <- iris3[,,"Setosa"]

# Үүссэн матрицийг датафрейм болгох
X <- as.data.frame(X)

# Махаланобисын зайн квадрат болон онцгой утга эсэхийг илтгэсэн
# нэмэлт багана бүхий датафрейм
rstatix::mahalanobis_distance(X)
```

4 Копула

Копула

Копула бол тухайн тархалт нь жигд тархалт байдаг олон хэмжээст хамтын тархалтын функц бөгөөд түүнийг хувьсагчдын холбоо хамаарлыг судлах, хамтын тархалт байгуулахад ашигладаг.

Тодорхойлолт 2 ($p = 2$ үед)**.** Дараах чанартай $C : [0,1]^2 \to [0,1]$ функцийг копула функц гэнэ. Үүнд:

1. $\forall u \in [0,1] : C(0,u) = C(u,0) = 0$
2. $\forall u \in [0,1] : C(u,1) = u$ ба $C(1,u) = u$
3. $u_1 \leq v_1$ ба $u_2 \leq v_2$ байх $\forall (u_1,u_2),(v_1,v_2) \in [0,1] \times [0,1] : C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \geq 0$

Дээрх нөхцөлийг хангах Гауссын копула, Архимедын копула зэрэг копула функцийн янз бүрийн бүл байдаг.

Копула

Теорем 2 (Sklar-ын теорем)**.** F нь F_{X_1} ба F_{X_2} тухайн тархалтуудтай хамтын тархалтын функц байг. Тэгвэл $x_1, x_2 \in \mathbb{R}$ бүрийн хувьд

$$F(x_1,x_2) = C\{F_{X_1}(x_1), F_{X_2}(x_2)\}$$

байх C функц оршин байна. Хэрэв F_{X_1} ба F_{X_2} тасралтгүй бол C цор ганц байна.

Гауссын копула

$$C_p(u,v) = \int_{-\infty}^{\Phi_1^{-1}(u)} \int_{-\infty}^{\Phi_2^{-1}(v)} f_\rho(x_1,x_2)dx_1dx_2$$

Энд f_ρ нь ρ корреляцтай хоёр хэмжээст хэвийн тархалтын нягтын функц, Φ_1 болон Φ_2 нь нэг хэмжээст стандарт хэвийн тархалтын функцүүд юм. Тухайн тохиолдолд $\rho = 0$ үед

$$C_0(u,v) = \int_{-\infty}^{\Phi_1^{-1}(u)} f_{X_1}(x_1)dx_1 \int_{-\infty}^{\Phi_2^{-1}(v)} f_{X_2}(x_2)dx_2 = uv$$

байна.

Лекц IV
Олон хэмжээст хэвийн тархалт II

1 Шугаман хувиргалт

Хувиргалт

Теорем 3. $X \sim N_p(\mu,\Sigma)$, $A_{(q\times p)}$, $c \in \mathbb{R}^q$, $q \le p$ бол $Y = AX + c$ санамсаргүй вектор q хэмжээст хэвийн тархалттай, өөрөөр хэлбэл,

$$Y \sim N_q(A\mu + c, A\Sigma A^T)$$

байна.

Дээрх теорем нь өмнөх лекцээр үзсэн хувиргалтын үндсэн теоремын өргөтгөл бөгөөд баталгааг нь бие даан үзнэ үү.

$X = (X_1,\dots,X_p)^T$ санамсаргүй векторыг дараах байдлаар хоёр дэд векторт хуваая.

$$X \begin{cases} \begin{matrix} X_1 \\ \vdots \\ X_r \end{matrix} \end{cases} X_1 \qquad \begin{cases} \begin{matrix} X_{r+1} \\ \vdots \\ X_p \end{matrix} \end{cases} X_2$$

Өөрөөр хэлбэл $X_1 = (X_1,\dots,X_r)^T$ ба $X_2 = (X_{r+1},\dots,X_p)^T$ гэе. Тэгвэл Σ ковариацийн матриц

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

блок матриц болно. Энд $\Sigma_{11} = \mathrm{cov}(X_1,X_1)$, $\Sigma_{22} = \mathrm{cov}(X_2,X_2)$, $\Sigma_{12} = \mathrm{cov}(X_1,X_2)$, $\Sigma_{21} = \Sigma_{12}^T = \mathrm{cov}(X_2,X_1)$ байна. Мөн $X_1 \sim N_r(\mu_1,\Sigma_{11})$ ба $X_2 \sim N_{p-r}(\mu_2,\Sigma_{22})$ байна.

Жишээ 13. $X = (X_1, X_2, X_3)^T \sim N_3\left(\begin{pmatrix}0\\0\\1\end{pmatrix}, \begin{pmatrix}1 & 2 & 1\\2 & 5 & 2\\1 & 2 & 2\end{pmatrix}\right)$ санамсаргүй векторын $X_1 = \begin{pmatrix}X_1\\X_2\end{pmatrix}$ болон $X_2 = (X_3)$ дэд векторын тархалтын параметрийг ол.

Энэ тохиолдолд $r = 2$ байна. Ковариацийн матриц нь дараах блок матриц болно.

$$\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} = \left(\begin{array}{cc|c} 1 & 2 & 1\\ 2 & 5 & 2\\ \hline 1 & 2 & 2\end{array}\right)$$

$$\Sigma_{11} = \operatorname{cov}(X_1, X_1) = \begin{pmatrix}1 & 2\\2 & 5\end{pmatrix} \quad \Sigma_{12} = \operatorname{cov}(X_1, X_2) = \begin{pmatrix}1\\2\end{pmatrix}$$
$$\Sigma_{21} = \operatorname{cov}(X_2, X_1) = \begin{pmatrix}1 & 2\end{pmatrix} \quad \Sigma_{22} = \operatorname{cov}(X_2, X_2) = \begin{pmatrix}2\end{pmatrix}$$

Мөн $\mu_1 = (0, 0)^T$, $\mu_2 = (1)$ болно.

Теорем 4. $X = \begin{pmatrix}X_1\\X_2\end{pmatrix} \sim N_p(\mu, \Sigma)$, $X_1 \in \mathbb{R}^r$, $X_2 \in \mathbb{R}^{p-r}$ байг.

$$\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

бас

$$X_{1.2} = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$$

гэе. Тэгвэл $X_{1.2}$ ба X_2 хамааралгүй бөгөөд

$$X_{1.2} \sim N_r(\mu_{1.2}, \Sigma_{11.2}), \qquad X_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$$

байна. Энд $\mu_{1.2} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$, $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Баталгаа

$$X_2 = \underbrace{\begin{pmatrix}0 & I_{p-r}\end{pmatrix}}_{A}\begin{pmatrix}X_1\\X_2\end{pmatrix} = AX$$
$$X_{1.2} = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 = \underbrace{\begin{pmatrix}I_r & -\Sigma_{12}\Sigma_{22}^{-1}\end{pmatrix}}_{B}\begin{pmatrix}X_1\\X_2\end{pmatrix} = BX$$

буюу X_2 болон $X_{1.2}$ нь X векторын шугаман хувиргалт тул өмнөх теорем ёсоор $X_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$ бас $X_{1.2} \sim N_r(\mu_{1.2}, \Sigma_{11.2})$ болно. Энд

$$\mu_{1.2} = B\mu + 0 = \begin{pmatrix}I_r & -\Sigma_{12}\Sigma_{22}^{-1}\end{pmatrix}\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$$
$$\Sigma_{11.2} = B\Sigma B^T = \begin{pmatrix}I_r & -\Sigma_{12}\Sigma_{22}^{-1}\end{pmatrix}\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}I_r\\-\Sigma_{22}^{-1}\Sigma_{21}\end{pmatrix}$$
$$= \begin{pmatrix}\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0\end{pmatrix}\begin{pmatrix}I_r\\-\Sigma_{22}^{-1}\Sigma_{21}\end{pmatrix} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

байна. ...

Баталгааны үргэлжлэл $X_{1.2}$ болон X_2 санамсаргүй векторууд хамааралгүй болох нь

$$\operatorname{cov}(X_2, X_{1.2}) = \operatorname{cov}(AX, BX) = A\operatorname{cov}(X,X)B^T = A\Sigma B^T =$$

$$\left(\begin{array}{|ccc|}\hline 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \\ \hline\end{array}\begin{array}{|ccc|}\hline 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ \hline\end{array}\right)\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\left(\begin{array}{|ccc|}\hline 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ \hline\end{array}\atop (-\Sigma_{12}\Sigma_{22}^{-1})^T\right) =$$

$$\begin{pmatrix}\Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix} I_r \\ (-\Sigma_{12}\Sigma_{22}^{-1})^T\end{pmatrix} = \begin{pmatrix}\Sigma_{21} + \Sigma_{22}(-\Sigma_{12}\Sigma_{22}^{-1})^T\end{pmatrix} =$$

$$\begin{pmatrix}\Sigma_{21} - \Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21}{}^{11}\end{pmatrix} = 0_{p-r\times r}$$

□

2 Нөхцөлт тархалт

Нөхцөлт тархалт

Теорем 5. Өмнөх теоремын нөхцөл биелж байг. Тэгвэл $X_2 = x_2$ үеийн X_1 санамсаргүй векторын нөхцөлт тархалт нь

$$(X_1|X_2 = x_2) \sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

буюу $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ нөхцөлт математик дундаж болон $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ нөхцөлт ковариацийн матриц бүхий хэвийн тархалт байна.

Баталгаа Өмнөх теорем дахь $X_{1.2} = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ шугаман хувиргалтыг

$$X_1 = X_{1.2} + \Sigma_{12}\Sigma_{22}^{-1}X_2$$

байдлаар эргэн авч үзье. Энэ нь $X_2 = x_2$ үед $X_1 = X_{1.2} + \Sigma_{12}\Sigma_{22}^{-1}x_2$ болно. Түүнчлэн

$$X_{1.2} \sim N_r(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

тул X_1 нь

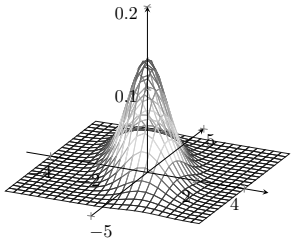
$$E(X_1|X_2 = x_2) = EX_{1.2} + \Sigma_{12}\Sigma_{22}^{-1}x_2 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\operatorname{cov}(X_1|X_2 = x_2) = \operatorname{cov}(X_{1.2} + \Sigma_{12}\Sigma_{22}^{-1}x_2) = \operatorname{cov}(X_{1.2}) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

нөхцөлт математик дундаж болон нөхцөлт ковариацийн матриц бүхий олон хэмжээст хэвийн тархалттай байна. □

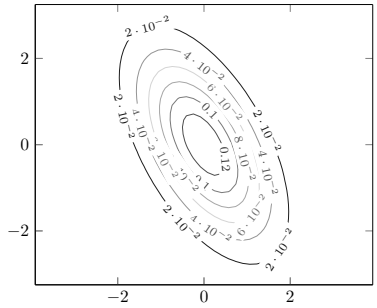
Жишээ 14. $X = \begin{pmatrix}X_1\\ X_2\end{pmatrix} \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & -0.8\\ -0.8 & 2\end{pmatrix}\right)$ бол X_2 санамсаргүй хувьсагчийн нөхцөл дэх X_1 санамсаргүй хувьсагчийн тархалтыг ол.

Жишээний хувьд $p = 2$, $r = 1$ байх тул $p - r = 1$, $\mu_1 = 0$, $\mu_2 = 0$, $\Sigma_{11} = 1$, $\Sigma_{12} = \Sigma_{21} = -0.8$, $\Sigma_{22} = 2$ ба нөхцөлт математик дундаж нь

$$E(X_1|X_2 = x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = -0.8 \cdot \frac{1}{2} \cdot x_2 = -0.4x_2$$



нягтын функцийн график



нягтын функцийн түвшний шугам

Зураг 8: X_1 ба X_2 санамсаргүй хувьсагчдын хамтын тархалт

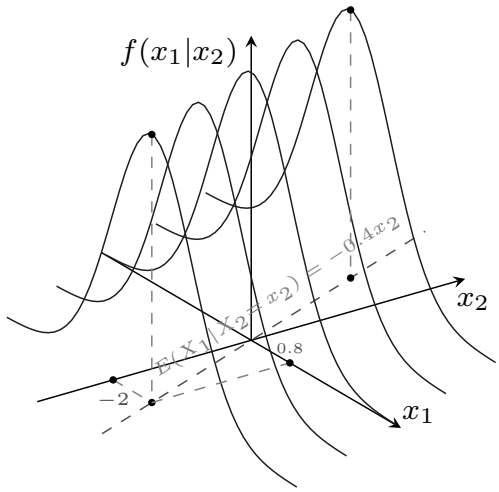
нөхцөлт ковариац нь

$$\text{cov}(X_1|X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - (-0.8)\frac{1}{2}(-0.8) = 0.68$$

гэж олдоно. Иймд $X_2 = x_2$ үеийн X_1 хувьсагчийн нөхцөлт тархалт нь $N_1(-0.4x_2, 0.68)$ буюу

$$f(x_1|x_2) = \frac{1}{\sqrt{2\pi 0.68}} \exp\left\{-\frac{(x_1 + 0.4x_2)^2}{2(0.68)}\right\}$$

гэж олдоно.



Зураг 9: $f(x_1|x_2) = \frac{1}{\sqrt{2\pi 0.68}} \exp\left\{-\frac{(x_1 + 0.4x_2)^2}{2(0.68)}\right\}$ нөхцөлт нягтын муруй

Нөхцөлт тархалт болон тухайн тархалтаар хамтын тархалт байгуулах

Теорем 6. $X_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$ ба $(X_1|X_2 = x_2) \sim N_r(Ax_2 + b, \Psi)$ бол

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_p(\mu, \Sigma)$$

болно. Энд

$$\mu = \begin{pmatrix} A\mu_2 + b \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Psi + A\Sigma_{22}A^T & A\Sigma_{22} \\ \Sigma_{22}A^T & \Sigma_{22} \end{pmatrix}$$

байна.

Жишээ 15. $X_2 \sim N_1(\mu_2 = 0, \Sigma_{22} = 1)$ ба $(X_1|X_2 = x_2) \sim N_2\left(Ax_2 + b = \begin{pmatrix} 2x_2 \\ x_2 + 1 \end{pmatrix}, \Psi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right.$ бол эдгээр хувьсагчдын хамтын тархалтыг ол.

Энэ тохиолдолд

$$Ax_2 + b = \begin{pmatrix} 2x_2 \\ x_2 + 1 \end{pmatrix} \Longrightarrow A = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ ба } b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

бас $\Psi = I_2$ тул

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_3\left(\mu = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}\right)$$

гэж олдоно.

3 Регрессийн шугаман загвар

Нөхцөлт математик дундаж ашигласан аппроксимац

$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $X_1 \in \mathbb{R}^r$, $X_2 \in \mathbb{R}^{p-r}$ гэе. Ингээд X_2 векторын тусламжтай X_1 векторын утгыг олох $X_1 = h(X_2) + U$ аппроксимац авч үзье.

Теорем 7. $E(X_1|X_2)$ нөхцөлт математик дундаж нь X_2 векторын тусламжтай X_1 векторыг аппроксимацлах $h(X_2) : \mathbb{R}^{p-r} \to \mathbb{R}^r$ функцүүд дундаас хамгийн бага дундаж квадрат алдаатай нь юм.

Дундаж квадрат алдааг $\text{MSE} = E\left\{(X_1 - h(X_2))^T(X_1 - h(X_2))\right\}$ гэж тодорхойлдог. Ийнхүү хэрэв алдааг дундаж квадрат алдаагаар хэмжинэ гэвэл нөхцөлт математик дундаж ашигласан

$$X_1 = E(X_1|X_2) + U$$

аппроксимац л хамгийн сайн нь юм.

Баталгаа $\{X_2 = x_2\}$ нөхцөлд $h(X_2)$ функц ердийн тогтмол болно. Иймд $\text{MSE} = E\left\{(X_1 - h(X_2))^T(X_1 - h(X_2))\right\}$ дундаж квадрат алдааг минимумчлах бодлого нь $\text{MSE} = E\left\{(X_1 - c)^T(X_1 - c)\right\}$ алдааг хамгийн бага болгох c тогтмолыг олох гэсэн бодлого руу шилжинэ. Үүнийг c аргументын хувьд дифференциалчлаад тэгтэй тэнцүүлбэл $E\{2(X_1 - c)\} = 0$ тэгшитгэл зохиогдох бөгөөд шийд нь $c = E(X_1)$ болно. Эцэст нь $\{X_2 = x_2\}$ нөхцөл тавьсанаа анхаарвал $h(X_2) = E(X_1|X_2)$ болно. □

Нөхцөлт математик дундаж ашигласан аппроксимацийн зарим чанар

$X_1 = E(X_1|X_2) + U$ аппроксимац нь дараах чанартай.

Чанар 8. 1. $EU = 0$

2. $E(U|X_2) = 0$

3. $\text{cov}(E(X_1|X_2), U) = 0$

Регрессийн шугаман загвар

X_1 болон X_2 санамсаргүй векторууд хамтдаа олон хэмжээст хэвийн тархалттай бол

$$\begin{aligned} X_1 &= E(X_1|X_2) + U \\ &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) + U \\ &= \beta_1 + BX_2 + U \end{aligned}$$

болно. Энд $B = \Sigma_{12}\Sigma_{22}^{-1}$, $\beta_1 = \mu_1 - B\mu_2$,

$$U \sim N_r(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

байна.

Энэ хэсгийг дуустал X_1 дэд векторын хэмжээсийг $r = 1$ гэе. Тэгвэл

$$X_1 = \beta_1 + BX_2 + U$$

шугаман загвар нь

$$X_1 = \beta_1 + \beta^T X_2 + U \qquad \text{энд } \beta^T = B_{(1\times p-1)} \text{ мөр вектор}$$

буюу

$$X_1 = \beta_1 + \beta_2 X_2 + \ldots + \beta_p X_p + U$$

хэлбэртэй болно.

Жишээ 16. $X = (X_1, X_2, X_3)$ буюу $p = 3$ үед $X_1 = (X_1)$ ба $X_2 = (X_2, X_3)$ болж улмаар

$$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

загвар бичигдэнэ.

Регрессийн шугаман загварын зүгээс харвал X_1 хувьсагчийн дисперс дараах байдлаар задарна.

$$\underbrace{D(X_1)}_{\text{нийт дисперс}} = \underbrace{D(\beta_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}_{\text{тайлбарлагдах дисперс}} + \underbrace{D(U)}_{\text{үл тайлбарлагдах дисперс}}$$

Регрессийн шугаман загварын хамааран хувьсагчийн дисперсэд эзлэх загвараар тайлбарлах дисперсийн хувийг тус загварын *детерминацийн коэффициент* гэдэг.

$$\rho^2 = \frac{D(\beta_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}{D(X_1)} = \frac{\text{cov}(\beta_1 + \beta^T X_2)}{D(X_1)} = \frac{\beta^T \text{cov}(X_2)\beta}{D(X_1)} = \frac{\sigma_{12}\Sigma_{22}^{-1}\sigma_{21}}{\sigma_{11}}$$

болно. Энд $\sigma_{11} = DX_1 = \Sigma_{11}$ нь скаляр, $\sigma_{12} = \Sigma_{12}$ нь $p-1$ ширхэг компоненттой мөр вектор, $\sigma_{21} = \Sigma_{21}$ нь $p-1$ ширхэг компоненттой багана вектор байна.

Жишээ 17. $\Sigma = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix}$ бол $X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U$ шугаман загварын детерминацийн коэффициентыг ол.

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{c|cc} 1 & 2 & 1 \\ \hline 2 & 5 & 2 \\ 1 & 2 & 2 \end{array}\right)$$

буюу $\sigma_{11} = 1$, $\sigma_{12} = \begin{pmatrix} 2 & 1 \end{pmatrix}$, $\sigma_{21} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $\Sigma_{22} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$ болж улмаар

$$\rho^2 = \frac{\sigma_{12}\Sigma_{22}^{-1}\sigma_{21}}{\sigma_{11}} = \frac{5}{6} \approx 0.833$$

утга олдоно.

```
# Ковариацийн матриц
Sigma <- matrix(c(1,2,1,2,5,2,1,2,2), "ncol" = 3, byrow = TRUE)

# Санамсаргүй өгөгдөл
set.seed(0)
X <- MASS::mvrnorm(n = 10000, mu = rep.int(x = 0, times =
    ncol(Sigma)), Sigma = Sigma)
X <- as.data.frame(X)

# Шугаман загвар
fit <- lm(formula = V1 ~ ., data = X)

# Детерминацийн коэффициент
summary(fit)$r.squared
```

```
0.8277409
```

Регрессийн шугаман загварын тайлбарлах хувьсагчдын өмнөх коэффициент сая үзсэнчлэн ерөнхий тохиолдолд $B = \Sigma_{12}\Sigma_{22}^{-1}$ харин $r = 1$ тухайн тохиолдолд $\beta^T = \sigma_{12}\Sigma_{22}^{-1}$ байна.

Жишээ 18. Өмнөх жишээний хувьд β_2 болон β_3 коэффициентуудыг олъё.

$\sigma_{12} = \begin{pmatrix} 2 & 1 \end{pmatrix}$, $\Sigma_{22}^{-1} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$ тул $\beta^T = \sigma_{12}\Sigma_{22}^{-1} = \begin{pmatrix} 1/3 & 1/6 \end{pmatrix} \approx \begin{pmatrix} 0.333 & 0.167 \end{pmatrix}$ хариу гарна.

```
coefficients(fit)[-1]
```

```
       V2        V3
0.3343419 0.1626271
```

4 Нөхцөлт болон тухайн корреляц

Нөхцөлт корреляц

Өмнөх теоремуудын томьёолол болон баталгаанд

$$X_1 = X_{1.2} + \Sigma_{12}\Sigma_{22}^{-1}X_2$$

хэлбэртэй шугаман илэрхийлэл болон холбогдох үр дүнгүүд дурдагдсан. Тэдгээрээс дараах дүгнэлт гарна. $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ санамсаргүй векторын хувьд X_1 буюу хамааран хувьсагч нь X_2 буюу үл хамааран хувьсагчаас шугаманаар хамаардаг бол $X_2 = x_2$ үеийн X_1 санамсаргүй векторын нөхцөлт ковариацийн матриц нь

$$\Sigma_{X_1|X_2=x_2} = \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

байна. Иймд энэхүү нөхцөлт ковариацийн матрицаас олдох корреляцийн коэффициент нь $X_2 = x_2$ үеийн X_1 вектор дахь хувьсагчдын нөхцөлт корреляцийг илэрхийлнэ.

Нөхцөлт корреляц ба тухайн корреляц

Түүнчлэн нөхцөлт ковариацтай зэрэгцэн

$$E(X_1|X_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = (\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2) + \Sigma_{12}\Sigma_{22}^{-1}x_2$$

үр дүнд хүрсэн. Үгтаа энэ нь $X_1 = (\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2) + \Sigma_{12}\Sigma_{22}^{-1}x_2 + U_1$ шугаман загвар тодорхойлно. Мөн хоёр санамсаргүй хувьсагч хоорондын тухайн ковариацийг олохын тулд хийдэг, тэдгээр хувьсагч дээрх хөндлөнгийн нийтлэг нөлөөг тооцож зайлуулах үйлдэлд шугаман загвар ашигладаг. Иймд олон хэмжээст хэвийн тархалттай санамсаргүй хувьсагчдын нөхцөлт корреляц ба тухайн корреляц хоёр эквивалент байна. Ерөнхийдөө нөхцөлт математик дундаж нь шугаман хэлбэртэй байдаггүйг санавал энэхүү эквивалент чанар тархалт бүрийн хувьд хүчинтэй байх албагүй юм.

Нөхцөлт ковариац ба тухайн ковариац хоёр эквивалент байх нөхцөл

Теорем 8. X_1 ба X_2 санамсаргүй векторуудын хувьд дараах хоёр нөхцөл эквивалент юм.

1. $E(X_1|X_2) = \beta + BX_2$ энд β нь вектор, B нь матриц
2. $\Sigma_{X_1||X_2} = E(\Sigma_{X_1|X_2})$

Олон хэмжээст хэвийн тархалтын хувьд нөхцөлт математик дунджийн илэрхийлэл нь шугаман хэлбэртэй байдаг. Бас нөхцөлт ковариацийн матриц нь тогтмол байдаг. Ингэхээр дээрх теоремыг батлах нь олон хэмжээст тархалтын хувьд түүний тухайн корреляц ба нөхцөлт корреляц хоёр тэнцүү гэдгийг харуулсан явдал болно.

Баталгаа Бүтэн ковариацын томьёо ба ковариацийн чанар ёсоор

$$\begin{aligned}\Sigma_{X_1||X_2} = \text{cov}(X_1 - h(X_2)) &= E(\text{cov}(X_1 - h(X_2)|X_2)) + \text{cov}(E(X_1 - h(X_2)|X_2)) \\ &= E(\text{cov}(X_1|X_2)) + \text{cov}(E(X_1 - h(X_2)|X_2)) \\ &= E(\Sigma_{X_1|X_2}) + \text{cov}(E(X_1 - h(X_2)|X_2))\end{aligned}$$

болно. Эндээс

$$\Sigma_{X_1||X_2} = E(\Sigma_{X_1|X_2}) \Leftrightarrow \text{cov}(E(X_1 - h(X_2)|X_2)) = 0 \Leftrightarrow P(E(X_1 - h(X_2)|X_2) = \beta) = 1$$

цувраа дүгнэлт гарна. Энд β нь тогтмол вектор юм. Математик дунджийн чанар ёсоор $E(X_1|X_2) = \beta + h(X_2)$ болох ба тухайн ковариацын хувьд $h()$ функцийг шугаманаар авдаг тул уг нөхцөлт математик дундаж нь $E(X_1|X_2) = \beta + BX_2$ хэлбэртэй болно. □

Жишээ 19.

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & -1 \\ 1 & 2 & 2 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \right)$$

санамсаргүй вектор авч үзье. X_1 ба X_2 хувьсагч тус бүрээс X_3 болон X_4 хувьсагчдын нөлөөг зайлуулсаны дараах тэдгээрийн хоорондох тухайн корреляцийг ол.

Эдгээр дөрвөн хувьсагч хамтдаа олон хэмжээст хэвийн тархалттай бөгөөд уг тархалтын хувьд тухайн корреляц нь нөхцөлт корреляцтай тэнцүү байдаг. Иймд $\rho_{X_1,X_2|X_3,X_4}$ нөхцөлт корреляц олно.

X санамсаргүй векторыг $X_1 = (X_1, X_2)^T$ ба $X_2 = (X_3, X_4)^T$ хоёр дэд векторт хуваавал ковариацийн матриц нь дараах байдалтай блок матриц болно.

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & -1 \\ \hline 1 & 2 & 2 & -1 \\ 0 & -1 & -1 & 4 \end{array} \right)$$

Тэгвэл

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \frac{3}{7} & 1 \\ 1 & 3 \end{pmatrix}$$

болох тул X_3 ба X_4 хувьсагчдын нөхцөл дэх X_1 ба X_2 хоёр санамсаргүй хувьсагчийн нөхцөлт корреляц

$$\rho_{X_1,X_2|X_3,X_4} = \frac{1}{\sqrt{3/7}\sqrt{3}} = \frac{\sqrt{7}}{3} \approx 0.882$$

гэж олдоно.

Ковариацийн матрицын урвуу ашиглаж тухайн корреляц олох

Тухайн корреляцийг Σ ковариацийн матрицын урвуугийн тусламжтай олж болно. $P = \Sigma^{-1}$ гэе. Үүнийг мөн $P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$ гэж бичвэл блок матрицын урвуугийн

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} \left(A - BD^{-1}C\right)^{-1} & -\left(A - BD^{-1}C\right)^{-1} BD^{-1} \\ -D^{-1}C\left(A - BD^{-1}C\right)^{-1} & D^{-1} + D^{-1}C\left(A - BD^{-1}C\right)^{-1} BD^{-1} \end{pmatrix}$$

адилтгал ёсоор $P_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ харьцаа гарна. Иймд X_1 дэд вектор дахь i болон j дүгээр хувьсагчдын хоорондох тухайн (i болон j дүгээр хувьсагчдаас бусад хувьсагчдын нөхцөл дэх) корррреляцийг P_{11}^{-1} матрицаас

$$\rho_{X_i,X_j|X\setminus\{X_i,X_j\}} = \frac{{P_{11}^{-1}}_{ij}}{\sqrt{{P_{11}^{-1}}_{ii}{P_{11}^{-1}}_{jj}}}$$

байдлаар олно. Энд ${P_{11}^{-1}}_{ij}$ бол P_{11}^{-1} матрицын i дүгээр мөр ба j дүгээр баганын огтлолцол дахь элемент юм.

Цаашилбал 2×2 матрицын урвуугийн

$$A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{|A|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

томьёо ёсоор

$$\rho_{X_i,X_j|X\setminus\{X_i,X_j\}} = \frac{{P_{11}^{-1}}_{ij}}{\sqrt{{P_{11}^{-1}}_{ii}{P_{11}^{-1}}_{jj}}} = \frac{-\frac{1}{|P_{11}|}P_{11ij}}{\sqrt{\frac{1}{|P_{11}|}P_{11jj}\frac{1}{|P_{11}|}P_{11ii}}} = -\frac{P_{11ij}}{\sqrt{P_{11jj}P_{11ii}}}$$

томьёо гарна. Иймд хамтдаа олон хэмжээст хэвийн тархалттай $X_1, \ldots, X_p$ санамсаргүй хувьсагчдын i болон j дүгээр хувьсагчдын хоорондох тухайн (i болон j дүгээр хувьсагчдаас бусад хувьсагчдын нөхцөл дэх) корррреляцийг олохдоо $X_1, \ldots, X_p$ хувьсагчдын ковариацийн матрицын урвуугаас шууд олсон корреляцийн коэффициентыг эсрэг тэмдэгтэйгээр авна.

Жишээ 20. Өмнөх жишээнд олсон корреляцийг ковариацийн матрицын урвуу ашиглаж ол.

$$\Sigma^{-1} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & -1 \\ 1 & 2 & 2 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}^{-1} = \frac{1}{2}\begin{pmatrix} 21 & -7 & -5 & -3 \\ -7 & 3 & 1 & 1 \\ -5 & 1 & 3 & 1 \\ -3 & 1 & 1 & 1 \end{pmatrix}$$

$$\rho_{X_1,X_2|X_3,X_4} = -\frac{-\frac{7}{2}}{\sqrt{3/2}\sqrt{21/2}} = \frac{\sqrt{7}}{3} \approx 0.882$$

R програмаар түүврийн тухайн корреляц ба хагас тухайн корреляц олох

Түүврийн тухайн корреляц болон хагас тухайн корреляц олоход ашиглаж болох багцын нэг бол `ppcor` юм. Мөн уг багц дахь функцүүдэд тухайн корреляц тэгтэй тэнцүү эсэх тухай таамаглал шалгах шинжүүрийг тусгасан байдаг.

Жишээ 21. R програмын `datasets` багцад бэлэн байдаг, цахилдаг цэцгийн цоморлиг болон дэлбээний урт ба өргөний хэмжээг илэрхийлсэн өгөгдөл бүхий `iris` датафреймаас Virginica зүйлд холбогдох мэдээллийг ялган авч улмаар дэлбээний урт ба цоморлигийн өргөн нь дэлбээний өргөн ба цоморлигийн уртын нөлөөг тооцсон үед хамааралгүй болохыг шалга.

Заасан өгөгдлийг дараах байдлаар ялгаж авна.

```
X <- subset("x" = iris, "subset" = Species == "virginica",
    "select" = 1:4)
```

Дэлбээний урт ба цоморлигийн өргөн хамааралгүй гэсэн таамаглалыг Пирсоны корреляцийн шинжүүрээр шалгахад няцааж (p-утга ≈ 0.0039) буйг дараах кодыг ажиллуулж үзээд харж болно.

```
cor.test(x = X$Petal.Length, y = X$Sepal.Width)
```

Харин дэлбээний өргөн ба цоморлигийн уртын нөлөөг зайлуулсаны дараа дэлбээний урт ба цоморлигийн өргөн хамааралгүй гэсэн таамаглалыг шалгахын тулд дараах байдалтай код бичиж ажиллуулна.

```
ppcor::pcor.test(x = X$Petal.Length, y = X$Sepal.Width, z =
    X[c("Petal.Width", "Sepal.Length")])
```

Эндээс p-утга ≈ 0.6096 магадлалын утга гарах бөгөөд энэ нь тухайн корреляц тэгтэй гэсэн таамаглалыг няцаах үндэслэлгүйг илтгэнэ. Иймд цахилдаг цэцгийн Virginica зүйлийн дэлбээний урт ба цоморлигийн өргөн хоорондын холбоо хамаарал нь дэлбээний өргөн ба цоморлигийн уртын нөлөөнөөс л үүдэлтэй гэсэн дүгнэлт гаргаж болох юм.

Түүнчлэн өгсөн датафрейм дахь хувьсагчдын хослол бүрийн хувьд бусад хувьсагчдын нөлөөг зайлуулсан тухайн корреляцийн матриц ба тухайн корреляц тэгтэй тэнцүү гэсэн тэг таамаглал шалгах шинжүүрийн магадлалын утга болон шинжүүрийн статистикийн туршилтын утга зэрэг нийлмэл үр дүнг нэг дор олдог `pcor()` гэдэг функц тус `ppcor` багцад байдаг. Жишээ болгон тус функцийн буцаах утгын `estimate` элемент буюу тухайн корреляцийн матрицыг шууд дуудан гаргав.

```
ppcor::pcor(X)$estimate
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  0.26908423   0.83815645  -0.1253643
Sepal.Width     0.2690842  1.00000000  -0.07558634   0.4838714
Petal.Length    0.8381564 -0.07558634   1.00000000   0.1796604
Petal.Width    -0.1253643  0.48387140   0.17966044   1.0000000
```

Хагас тухайн корреляцын коэффициентыг R програмын `ppcor` нэмэлт багц дахь `spcor()` болон `spcor.test()` функцийн тусламжтай олж болно. Эдгээр

функцүүд нь хөндлөнгийн ”гуравдагч” хувьсагчийн нөлөөг хоёр дахь хувьсагчаас зайлуулж, эхний хувьсагчийг хэвээр үлдээн тус корреляцыг тооцоолдог.

Лекц V
Олон хэмжээст хэвийн тархалт III

1 Параметрийн үнэлэлт

Олон хэмжээст түүвэр

Параметрийн утгыг үнэлж олоход өгөгдөл зайлшгүй шаардлагатай. $X = (X_1, \ldots, X_p)^T \sim N_p(\mu, \Sigma)$ санамсаргүй векторын эх олонлогоос авсан n хэмжээтэй түүврийг $\mathcal{X}$ гэж тэмдэглэнэ. Түүврийн i дүгээр элемент

$$x_i = (x_{i,1}, \ldots, x_{i,p})^T$$

хэлбэртэй байна.

$$\mathcal{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix}$$

Ингээд дээрх түүврийг хэрэгжүүлж өгөгдөл цугуулсан гэж үзнэ.

Өгөгдлийн жишээ

```
datasets::mtcars |> head()
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear
    carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4
     4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4
     4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4
     1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3
     1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3
     2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3
     1
```

Энэ нь $\mathcal{X}$ матрицын бүтцийг харуулах зорилготой жишээ бөгөөд олон хэмжээст хэвийн тархалттай уялдаагүй юм.

Түүврийн дундаж ба түүврийн ковариацын матриц

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Энд n нь түүврийн хэмжээ, x_i нь түүврийн i дүгээр элемент буюу $\mathcal{X}$ түүврийн матрицын i дүгээр мөр юм.

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T$$

μ параметрийн үнэлэлт

Параметрийн үнэлэлт олоход хэрэглэдэг моментын аргын дагуу түүврийн дундаж нь "онолын" дунджийн үнэлэлт болно.

Дунджийн шугаман чанар бас түүврийг $X \sim N_p(\mu, \Sigma)$ санамсаргүй векторын эх олонлогоос авсан бөгөөд тус санамсаргүй векторын дундаж нь μ параметрын утгатай тэнцүү зэргийг анхаарвал

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

буюу түүврийн дундаж нь олон хэмжээст хэвийн тархалтын дундаж утгын вектор μ параметрийн хазайлтгүй үнэлэлт болно.

Σ параметрийн үнэлэлт

Уг параметрийг $\hat{\Sigma} = S$ гэж үнэлнэ. Хазайлтгүй үнэлэлт болохыг нь дараах байдлаар шалгана.

$$\begin{aligned} S &= \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \mu) - (\bar{X} - \mu)][(x_i - \mu) - (\bar{X} - \mu)]^T \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T - n(\bar{X} - \mu)(\bar{X} - \mu)^T\right] \end{aligned}$$

Одоо $\bar{X} \sim N_p(\mu, \Sigma/n)$ болохыг анхаарвал

$$E(S) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(x_i - \mu)(x_i - \mu)^T - nE(\bar{X} - \mu)(\bar{X} - \mu)^T\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}\Sigma - n\frac{\Sigma}{n}\right] = \Sigma$$

байна. Энэ бол S түүврийн ковариацын матриц нь Σ ковариацын матрицын хазайлтгүй үнэлэлт гэсэн үг юм.

2 Олон хэмжээст хэвийн тархалттай эсэх тухай таамаглал

Олон хэмжээст хэвийн тархалтын Shapiro–Wilk-ийн шинжүүр

Шинжүүрийн статистик $X \sim N_p(\mu, \Sigma)$ санамсаргүй векторыг $Z = \Sigma^{-1/2}(X - \mu)$ Махаланобисын хувиргалтаар $Z = (Z_1, \ldots, Z_p)$ гэсэн хамааралгүй, стандарт

хэвийн тархалттай хувьсагчдаас тогтох вектор руу шилжүүлдэг хэмээн өмнө үзсэн. Энд $\Sigma^{-1/2}$ нь ковариацийн матрицын урвуугийн симметр бөгөөд эерэг тодорхойлогдсон квадрат язгуур юм. Өгөгдөл дээр тус хувиргалтыг $\hat{\mu} = \bar{X}$ болон $\hat{\Sigma} = S$ үнэлэлтүүдийн тусламжтай хэрэгжүүлэх боломжтой. Ийнхүү холбоо хамааралтай хувьсагчдыг хамааралгүй буюу саланги тусдаа байдал руу шилжүүлж чадна. Иймд

$$H_0 : X \sim N_p(\mu, \Sigma), \quad \mu \text{ болон } \Sigma \text{ мэдэгдэхгүй}$$

тэг таамаглалыг дээрх хувиргалтаар үүсэх $Z_1, \ldots, Z_p$ хувьсагч тус бүр дээрх Shapiro–Wilk-ийн шинжүүрийн W_{Z_i} статистикуудаас тогтох

$$W^* = \frac{1}{p} \sum_{i=1}^{p} W_{Z_i}$$

статистикийн тусламжтай шалгаж болно.

Шинжүүрийн няцаах муж H_0 үнэн үед W_{Z_i} бүр нэгд ойр байдаг тул W^* статистик ч бас нэгд ойр утгатай байна. Харин W^* статистикийн утга нь шинжүүрийн няцаах утгаас бага бол тэг таамаглалыг няцаана. Өөрөөр хэлбэл уг шинжүүрийн няцаах муж $\{W^* < c_\alpha\}$ хэлбэртэй байна. α итгэх түвшинтэй шинжүүрийн няцаах утга c_α нь

$$\alpha = P(W^* < c_\alpha | H_0)$$

тэгшитгэлээр тодорхойлогдоно. c_α няцаах утгыг олоход тэг таамаглал үнэн гэсэн нөхцөл дэх W^* статистикийн тархалт шаардлагатай. Гэвч W^* статистикийг зохиоход оролцсон нэг хэмжээст хэвийн тархалт дээрх Shapiro–Wilk-ийн шинжүүрийн W_{Z_i} статистикийн тархалт аналитик аргаар олддоггүй учраас тус шинжүүрийн статистикийн тархалт ч аналитик аргаар олдохгүй. Иймд нэг хэмжээст хэвийн тархалт дээрх Shapiro–Wilk-ийн шинжүүрийг олон хэмжээст хэвийн тархалтын хувьд ийнхүү өргөтгөсөн судлаачид нь өмнөх шинжүүрийн няцаах утгыг стандарт хэвийн тархалтын хувиргалтын тусламжтай ойролцоогоор олсонтой адил төстэй арга санал болгосон. Мөн энэхүү ойролцоо утга нь түүврийн хэмжээ n өсөхөд W^* статистикийн тэг тархалтынхтай улам ойролцоо болдогийг симуляцийн туршилтаар тогтоожээ.

R програм дээрх бэлэн функц Олон хэмжээст хэвийн тархалт дээрх Shapiro–Wilk-ийн шинжүүрийг `rstatix` багц дахь `mshapiro_test()` функцийн тусламжтай ашиглаж болно.

Жишээ 22. R програмын `datasets` багцад байдаг, цахилдаг цэцгийн тухай мэдээлэл агуулсан `iris3` массив авч үзье. Setosa зүйлийн хувьд цоморлигийн урт ба өргөн нь хамтдаа олон хэмжээст хэвийн тархалттай болохыг Shapiro–Wilk-ийн шинжүүрээр 0.05 итгэх түвшинд шалга.

```
rstatix::mshapiro_test(datasets::iris3[,c("Sepal L.","Sepal
    W."),"Setosa"])
```

```
 statistic p.value
     0.973   0.302
```

Шинжүүрийн магадлалын утга 0.302 гэж олдсон нь 0.05 итгэх түвшингээс их тул тэг таамаглалыг үл няцаана.

3 Үнэний хувийн харьцаат шинжүүр

Үнэний хувийн харьцаат шинжүүр

Үл мэдэгдэх θ параметр бүхий $f_X(x,\theta)$ нягттай X санамсаргүй вектор авч үзнэ.

Таамаглал

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$$

Энд Θ_0 нь θ параметрийн талаар таамаглаж буй утгуудаас тогтох олонлог харин Θ нь θ параметрийн утгын олонлог юм.

Шинжүүрийн статистик

$$\lambda = \lambda(\mathcal{X}) = \frac{L_0}{L_.} = \frac{\max\limits_{\theta\in\Theta_0} L(\mathcal{X},\theta)}{\max\limits_{\theta\in\Theta} L(\mathcal{X},\theta)}$$

Энд L нь үнэний хувийн функц юм.

Шинжүүрийн няцаах муж H_0 үнэнд ойртоход L_0 нь $L_.$ рүү дөхөх тул λ харьцааны утга 1 рүү зүүнээс нь тэмүүлнэ. Эсрэг тохиолдолд холдоно. Иймд няцаах муж $\{\mathcal{X} : \lambda(\mathcal{X}) < c\}$ хэлбэртэй байна.

Шинжүүрийн асимптот няцаах муж α итгэх түвшинтэй асимптот няцаах муж Wilks-ийн теоремоор

$$\{\mathcal{X} : -2\ln\lambda(\mathcal{X}) > \chi^2_{\alpha,q-r}\}$$

хэлбэртэй байдаг. Энд r болон q нь харгалзан тэг таамаглалаар заагласан болон зааглаагүй байх үеийн үл мэдэгдэх параметрүүдийн тоо юм.

4 Параметрийн таамаглалууд

Дундаж утгын векторын тухай таамаглал

Таамаглал

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

Энд Σ мэдэгдэнэ гэж тооцно.

Шинжүүрийн статистик Таамаглалыг үнэний хувийн харьцаат шинжүүрээр шалгана. Үүний тулд

$$-2\ln\lambda = -2\ln\frac{L_0}{L_.} = -2\ln\frac{\max\limits_{\theta\in\Theta_0} L(\mathcal{X},\theta)}{\max\limits_{\theta\in\Theta} L(\mathcal{X},\theta)}$$

статистик дахь $\ln L_0$ болон $\ln L_.$ логарифм үнэн хувиудыг олох хэрэгтэй. Эхлээд олон хэмжээст хэвийн тархалтад харгалзах логарифм үнэний хувийн функцийн илэрхийллийг олно.

$\mathcal{X}$ түүвэр болон олон хэмжээст хэвийн тархалтын нягтын функцийг үнэний хувийн функцийн томьёонд орлуулаад эмхэтгэвэл

$$L(\mathcal{X},\mu) = f(x_1,\mu)\cdot\ldots\cdot f(x_n,\mu)$$

$$= |2\pi\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right\}$$

болох ба үүний логарифм

$$\ln L(\mathcal{X}, \mu) = -\frac{1}{2} \ln |2\pi\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

хэлбэртэй байна.

Сүүлийн илэрхийлэл дэх квадратлаг хэлбэрийг $\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$ чанар ашиглан хувиргавал

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) &= \sum_{i=1}^{n} (x_i - \bar{X})^T \Sigma^{-1} (x_i - \bar{X}) + n(\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu) \\
&= \sum_{i=1}^{n} \mathrm{tr}\left\{ (x_i - \bar{X})^T \Sigma^{-1} (x_i - \bar{X}) \right\} + n(\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu) \\
&= \mathrm{tr}\left\{ \Sigma^{-1} \sum_{i=1}^{n} (x_i - \bar{X})^T (x_i - \bar{X}) \right\} + n(\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu) \\
&= \mathrm{tr}\{\Sigma^{-1} n S\} + n(\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu)
\end{aligned}$$

цэгцтэй илэрхийлэлд хүрнэ. Үүнийг буцаан орлуулбал

$$\ln L(\mathcal{X}, \mu) = -\frac{1}{2} \ln |2\pi\Sigma| - \frac{n}{2} \mathrm{tr}\{\Sigma^{-1} S\} - \frac{n}{2} (\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu)$$

болно.

$\mu = \mu_0$ гэсэн тэг таамаглалаар зааглагдсан логарифм үнэний хувь

$$\ln L_0 = \ln L_0(\mathcal{X}, \mu_0) = -\frac{1}{2} \ln |2\pi\Sigma| - \frac{n}{2} \mathrm{tr}\{\Sigma^{-1} S\} - \frac{n}{2} (\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0)$$

болно. Харин зааглалтгүй логарифм үнэний хувь $\ln L_.$ хэмжигдэхүүний хувьд логарифм үнэний хувийн функцийг μ параметрээр максимумчлах ёстой. Үүнийг максимумчлах нь μ дундаж оролцсон, сөрөг утгатай сүүлийн гишүүнийг минимумчлахтай эквивалент юм. Тус илэрхийлэл $\mu = \bar{X}$ үед хамгийн бага 0 гэсэн утгадаа хүрнэ. Иймхүү

$$\ln L_. = -\frac{1}{2} \ln |2\pi\Sigma| - \frac{n}{2} \mathrm{tr}\{\Sigma^{-1} S\}$$

болно. Иймд шинжүүрийн статистик

$$-2 \ln \lambda = -2 \left[\ln L_0 - \ln L_. \right] = n(\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0)$$

хэлбэртэй болно.

Шинжүүрийн асимптот няцаах муж Шинжүүрийн няцаах мужийг эцэслэн заахад зөвхөн $-2 \ln \lambda$ статистикийн асимптот хи-квадрат тархалтын чөлөөний зэрэг л дутуу үлдсэн. Σ мэдэгдэнэ гэж тооцсоныг анхаарвал $H_0 : \mu = \mu_0$ тэг таамаглалаар зааглагдсан байх үеийн олон хэмжээст хэвийн тархалтын

үл мэдэгдэх параметрийн тоо 0 болно. Харин зааглалтгүй үед $\mu = (\mu_1, \ldots, \mu_p)$ гэсэн p ширхэг параметрийн утга мэдэгдэхгүй. Иймд тус хи-квадрат тархалтын чөлөөний зэрэг $q - r = p - 0 = p$ болно. Тэгэхээр дундаж утгын векторын тухай таамаглал шалгахад ашиглах α итгэх түвшинтэй шинжүүрийн асимптот няцаах муж

$$n(\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0) > \chi^2_{\alpha,p}$$

хэлбэртэй байна.

Дундаж утгын векторын тухай таамаглал

Таамаглал

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

Энд Σ мэдэгдэхгүй гэж тооцно.

Шинжүүрийн няцаах муж

Өмнөхтэй төстэй байдлаар

$$n \ln \left\{ 1 + (\overline{x} - \mu_0)^T S^{-1} (\overline{x} - \mu_0) \right\} > \chi^2_{\alpha,p}$$

хэлбэртэй шинжүүрийн муж олдоно. Мөн практикт

$$\frac{n-p}{p} \frac{n}{n-1} (\overline{x} - \mu_0)^T S^{-1} (\overline{x} - \mu_0) > F_{\alpha,p,n-p}$$

шинжүүрийн мужийг өргөн ашигладаг.

Жишээ 23. R програмын `datasets` багцын `iris3` массивт хадгалаатай байдаг цахилдаг цэцгийн цоморлиг болон дэлбээний хэмжээний өгөгдөл авч үзье. Түүний Setosa зүйлийн хувьд цоморлиг болон дэлбээний хэмжээ заасан дөрвөн хувьсагчийн дунджийн тухай

$$H_0 : \mu = (5.0, 3.4, 1.5, 0.2)$$

таамаглалыг F шинжүүрээр $\alpha = 0.05$ итгэх түвшинд шалга.

Өгөгдөл

```
X <- iris3[ , , "Setosa"]
```

Таамаглал

```
mu_0 <- c(5.0, 3.4, 1.5, 0.2)
```

Тооцоо

```
p <- ncol(X); n <- nrow(X); m <- colMeans(X); S <- cov(X)
# шинжүүрийн статистикийн туршилтын утга
as.numeric((n-p)/p * n/(n-1) * t(m - mu_0) %*% solve(S) %*% (m -
    mu_0))
# шинжүүрийн няцаах утга
qf(p = 0.05, df1 = p, df2 = n - p, lower.tail = FALSE)
```

$F \approx 4.030 > F_{\alpha=0.05,p=4,n-p=46} \approx 2.574$ тул тэг таамаглалыг няцаана.

Ашиглаж болох бэлэн функц

```
ICSNP::HotellingsT2(X = X, mu = mu_0, test = "f")
```

```
   Hotelling's one sample T2-test

data:  X
T.2 = 4.0302, df1 = 4, df2 = 46, p-value = 0.006953
alternative hypothesis: true location is not equal to
    c(5,3.4,1.5,0.2)
```

Ковариацийн матрицын тухай таамаглал

Таамаглал

$$H_0 : \Sigma = \Sigma_0, \quad H_1 : \Sigma \neq \Sigma_0$$

Энд μ мэдэгдэхгүй гэж тооцно.

Шинжүүрийн статистик

$$\ln L_0 = -\frac{n}{2}\ln|2\pi\Sigma_0| - \frac{n}{2}\operatorname{tr}(\Sigma_0^{-1}S) \qquad \ln L_1 = -\frac{n}{2}\ln|2\pi S| - \frac{np}{2}$$

тул шинжүүрийн статистик дараах хэлбэртэй болно.

$$-2\ln\lambda = 2(\ln L_1 - \ln L_0) = n\operatorname{tr}(\Sigma_0^{-1}S) - n\ln|\Sigma_0^{-1}S| - np$$

Шинжүүрийн няцаах муж

α итгэх түвшинтэй шинжүүрийн асимптот няцаах муж дараах байдалтай байна.

$$n\operatorname{tr}(\Sigma_0^{-1}S) - n\ln|\Sigma_0^{-1}S| - np > \chi^2_{\alpha,\frac{p(p+1)}{2}}$$

Жишээ 24. Өмнөх жишээнд авч үзсэн массиваас

```
X <- iris3[ , c("Sepal W.", "Petal L."), "Setosa"]
```

өгөгдөл ялгаж авъя. Sepal W. болон Petal L. хувьсагчдын хувьд

$$H_0 : \Sigma = \Sigma_0 = \begin{pmatrix} 0.135 & 0 \\ 0 & 0.031 \end{pmatrix}$$

таамаглалыг $\alpha = 0.05$ итгэх түвшинд шалга.

Таамаглал

```
Sigma_0 <- matrix(c(0.135, 0, 0, 0.031), ncol = 2)
```

Тооцоо

```
p <- ncol(X); n <- nrow(X); S <- cov(X)
# шинжүүрийн статистикийн туршилтын утга
n * psych::tr(solve(Sigma_0) %*% S) - n * log(det(solve(Sigma_0)
    %*% S)) - n * p
# шинжүүрийн няцаах утга
qchisq(p = 0.05, df = p*(p+1)/2, lower.tail = FALSE)
```

Дүгнэлт

$$-2\ln\lambda \approx 1.722 \;>\; \chi^2_{\alpha=0.05,df=3} \approx 7.815$$

харьцаа үл биелэх тул H_0 таамаглалыг $\alpha = 0.05$ итгэх түвшинд үл няцаана.

5 Дунджийн итгэх муж

Олон хэмжээст хэвийн тархалтын дунджийн итгэх муж

Ковариац нь мэдэгдэхгүй байх үед дунджийн тухай таамаглал шалгах F шинжүүрээс олон хэмжээст хэвийн тархалтын дундаж утгын векторын итгэх мужийн дараах хэлбэртэй томьёо гарна.

Томьёо 2 (Хэвийн тархалтын дунджийн $1-\alpha$ хувийн итгэх муж)**.**

$$\left\{\mu:\ (\mu-\overline{x})^T S^{-1}(\mu-\overline{x}) \le \frac{p}{n-p}F_{\alpha,p,n-p}\right\}$$

Жишээ 25. Өмнөх жишээнд авч үзэж байсан `iris3` массиваас ялган авсан

```
X <- as.data.frame(iris3[ , c("Petal L.", "Petal W."), "Setosa"])
```

өгөгдөл ашиглаж, цахилдагийн Setosa зүйлийн дэлбээний урт болон өргөний дундаж хэмжээний 95 хувийн итгэх завсар байгуул.

```
p <- ncol(X); print(p)
n <- nrow(X); print(n)
qf(p = 0.05, df1 = p, df2 = n - p, lower.tail = FALSE)
```

тушаалаар

```
2         # p
50        # n
3.190727  # F_{α=0.05,p=2,n−p=48}
```

үр дүн гарах тул дунджийн итгэх мужийн илэрхийллийн сүүлийн хэсэг

$$(\mu-\overline{x})^T S^{-1}(\mu-\overline{x}) \le \frac{p}{n-p}F_{\alpha,p,n-p} \approx \frac{2}{50-2}\cdot 3.190 \approx 0.133$$

болно.

Одоо $(\mu-\overline{x})^T S^{-1}(\mu-\overline{x}) \le 0.133$ илэрхийллийн эхний хэсгийг гаргая. Өмнө үзсэн хоёр хэмжээст хэвийн тархалтын нягтын илэрхийллээс

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]$$

томьёо гаргаж болно. Энд ρ нь корреляц юм. Иймд итгэх мужийн илэрхийлэл

$$\frac{1}{1-\rho^2}\left[\frac{(\mu_1-\overline{x}_1)^2}{\sigma_1^2} - \frac{2\rho(\mu_1-\overline{x}_1)(\mu_2-\overline{x}_2)}{\sigma_1\sigma_2} + \frac{(\mu_2-\overline{x}_2)^2}{\sigma_2^2}\right] \le 0.133$$

хэлбэртэй болно.

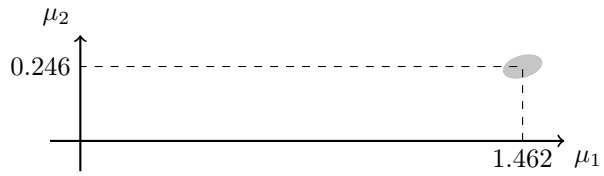
Өмнөх илэрхийлэл дэх статистикуудыг дараах байдлаар олно.

```
mean(X$‘Petal L.‘)  # x̄_1 = 1.462
mean(X$‘Petal W.‘)  # x̄_2 = 0.246
sd(X$‘Petal L.‘)    # σ_1 ≈ 0.174
sd(X$‘Petal W.‘)    # σ_2 ≈ 0.105
cor(x = X$‘Petal L.‘, y = X$‘Petal W.‘)  # ρ ≈ 0.332
```

Эдгээрийг орлуулбал дунджийн итгэх мужийн

$$\frac{(\mu_1 - 1.462)^2}{0.030} - \frac{(\mu_1 - 1.462)(\mu_2 - 0.246)}{0.027} + \frac{(\mu_2 - 0.246)^2}{0.011} \leq 0.118$$

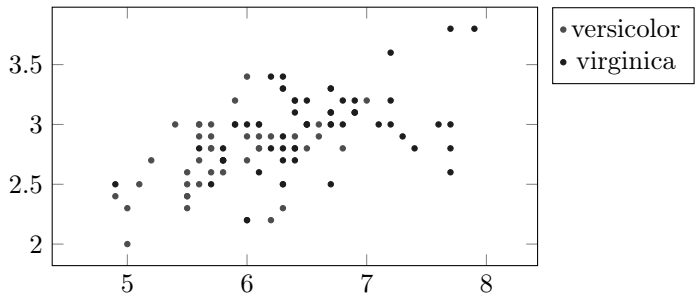
илэрхийлэл олдох бөгөөд дор зурж дүрслэв.



Лекц VI
Олон хэмжээст хэвийн тархалт IV

1 Дундаж утгын векторуудыг жиших

Дунджуудын зөрүүний тухай таамаглал



Зураг 10: Цахилдаг цэцгийн versicolor болон virginica зүйлүүдийн цоморлигийн урт болон өргөн

Жишээ 26. Цоморлигийн урт болон өргөний дундаж хэмжээ тус хоёр зүйлийн хувьд ялгаатай гэж батлах зорилго тавъя.

Өгөгдөл

Өгөгдөл ялгаж авах

```
X <- subset(
  "x" = iris,
  "subset" = Species %in% c("versicolor", "virginica"),
  "select" = c("Sepal.Length", "Sepal.Width", "Species")
)
```

Цэгэн диаграмм байгуулах

```
plot(
  x = X$Sepal.Length, y = X$Sepal.Width,
  xlab = "Sepal.Length", ylab = "Sepal.Width",
  asp = 1, pch = unclass(X$Species),
  col = c("red", "green", "blue")[unclass(X$Species)]
)
```

Ашиглах шинжүүр

Тус бүртээ $X_i \sim N_p(\mu_i, \Sigma)$ тархалттай X_1 болон X_2 санамсаргүй векторуудын эх олонлогоос харгалзан n_1 болон n_2 хэмжээтэй хамааралгүй түүврүүд авсан гэе. Энд Σ мэдэгдэхгүй гэж тооцно.

Таамаглал

$$H_0 : \mu_1 - \mu_2 = \Delta\mu, \quad H_1 : \mu_1 - \mu_2 \neq \Delta\mu$$

Энд $\Delta\mu$ нь таамаглаж буй утга юм.

Шинжүүрийн асимптот няцаах муж

$$\frac{n_1 n_2}{n_1 + n_2} \frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 2)} (\overline{x}_1 - \overline{x}_2 - \Delta\mu)^T S^{-1} (\overline{x}_1 - \overline{x}_2 - \Delta\mu) \geq F_{\alpha,p,n_1+n_2-p-1}$$

Энд $S = \dfrac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$ нь ковариацийн матрицуудын жинлэсэн дундаж буюу нийт түүврийн ковариацийн матриц юм.

```
delta.mu <- c(0, 0)

X1 <- X[X$Species == "versicolor", c("Sepal.Length",
    "Sepal.Width")]
X2 <- X[X$Species == "virginica", c("Sepal.Length",
    "Sepal.Width")]
p <- ncol(X1); n1 <- nrow(X1); n2 <- nrow(X2)
m1 <- colMeans(X1); m2 <- colMeans(X2)
S1 <- cov(X1); S2 <- cov(X2)
S <- {(n1-1)*S1 + (n2-1)*S2} / {n1+n2-2}
# шинжүүрийн статистикийн туршилтын утга
as.numeric({n1*n2}/{n1+n2} * {n1+n2-p-1}/{p*(n1+n2-2)} * t(m1 -
    m2 - delta.mu) %*% solve(S) %*% (m1 - m2 - delta.mu))
# шинжүүрийн няцаах утга
qf(p = 0.05, df1 = p, df2 = n1 + n2 - p - 1, lower.tail = FALSE)
```

Дунджуудын зөрүүний тухай уг таамаглалыг R програмын `ICSNP` багц дахь `HotellingsT2()` функцийн тусламжтай шалгах боломжтой.

```
ICSNP::HotellingsT2(X = X1, Y = X2, mu = delta.mu, test = "f")
```

`mu` аргументаар дамжуулж дундаж утгын векторуудын зөрүүний талаар таамаглаж буй утгаа өгнө.

```
   Hotelling's two sample T2-test

data:  X1 and X2
T.2 = 15.827, df1 = 2, df2 = 97, p-value = 1.126e-06
alternative hypothesis: true location difference is not equal to
    c(0,0)
```

Энэ нь бидний өмнөх тооцооны үр дүнтэй адил гарсан. Ийнхүү шинжүүрийн статистикийн туршилтын утга 15.827, няцаах утга 3.090 гарсан нь шинжүүрийн няцаах мужийн нөхцөлийг хангаж буй тул тэг таамаглалыг няцаана.

2 Ковариацийн матрицуудыг жиших

Ковариацүүд тэнцүү байх тухай таамаглал

Тус бүртээ $X_i \sim N_p(\mu_i, \Sigma_i)$ тархалттай k ширхэг $X_1, \ldots, X_k$ санамсаргүй векторуудын эх олонлогоос харгалзан $n_1, \ldots, n_k$ хэмжээтэй хамааралгүй түүвэрүүд авсан гэе.

Таамаглал

$$H_0 : \Sigma_1 = \ldots = \Sigma_k, \quad H_1 : \text{дор хаяж хоёр нь ялгаатай}$$

Жишээ 27. Өмнөх жишээнд дунджууд тэнцүү гэсэн таамаглал шалгахад ашигласан шинжүүр нь бүлгүүдийн ковариацийн матрицуудыг тэнцүү гэсэн нөхцөлтэй байсан бөгөөд бид тэрхүү нөхцөлийг шалгаж нягтлалгүйгээр тэрхүү шинжүүрийг ашигласан билээ. Иймд тэр нөхцөл биелэх эсэхийг шалгая.

Шинжүүрийн статистик

$$M = -2 \ln \lambda = n \ln |S| - \sum_{i=1}^{k} (n_i - 1) \ln |S_i|$$

Энд $S = \dfrac{(n_1 - 1)S_1 + \ldots + (n_k - 1)S_k}{n_1 + \ldots + n_k - k}$ байна.

Шинжүүрийн статистикийн асимптот тархалт

$$(1 - u)M \sim \chi^2_{\frac{1}{2}(k-1)p(p+1)}$$

Энд $u = \left(\sum\limits_{i=1}^{k} \dfrac{1}{n_i - 1} - \dfrac{1}{\sum\limits_{i=1}^{k} (n_i - 1)} \right) \dfrac{2p^2 + 3p - 1}{6(p+1)(k-1)}$ байна.

Ковариацүүд тэнцүү байх тухай таамаглал шалгахад зориулсан `biotools` багцын `boxM()` бас `rstatix` багцын `box_m()` зэрэг функц байдаг.

```
biotools::boxM(data, grouping)
```

Жишээний хувьд дараах код бичиж болно.

```
biotools::boxM(data = X[c("Sepal.Length", "Sepal.Width")],
    grouping = X$Species)
```

Эндээс дараах үр дүн гарна.

```
    Box's M-test for Homogeneity of Covariance Matrices

data:  X[c("Sepal.Length", "Sepal.Width")]
Chi-Sq (approx.) = 3.0878, df = 3, p-value = 0.3783
```

`p-value = 0.3783` буюу бидний сонгож авдаг 0.05 итгэх түвшнгээс их байгаа тул бүлгүүдийн ковариацийн матрицуудыг ялгаатай гэх үндэслэлгүй ажээ.

3 Олон хэмжээст дисперсийн шинжилгээ

Дунджууд тэнцүү байх тухай таамаглал буюу Олон хэмжээст дисперсийн шинжилгээ (MANOVA)

Нэг ижил ковариацийн матрицтай санамсаргүй векторын эх олонлогийн хэд хэдэн бүлгийг дундаж утгын вектороор нь харьцуулах шинжүүр үзье. Өөрөөр хэлбэл дараах санамсаргүй вектор болон таамаглал авч үзье.

$$X_i \sim N_p(\mu_i, \Sigma), \quad (i = 1, \ldots, k)$$

Таамаглал

$$H_0 : \mu_1 = \ldots = \mu_k, \quad H_1 : \text{ядаж хоёр нь ялгаатай}$$

Энд Σ ковариацийн матрицыг мэдэгдэхгүй гэж тооцно.

Шинжүүрийн статистик

$$\Lambda = \frac{|W|}{|B+W|} = \prod_{i=1}^{r} \frac{1}{1+\lambda_i}$$

Энд

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)^T \quad B = \sum_{i=1}^{k} (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T$$

бол $\lambda_1, \ldots, \lambda_r$ нь $W^{-1}B$ матрицын хувийн утгууд, r нь B матрицын ранг бөгөөд $r = \min\{p, k-1\}$ байна. Мөн энд n_i нь i дүгээр бүлгийн хэмжээ, x_{ij} нь i дүгээр бүлгээс авсан түүврийн j дүгээр элемент, $\overline{x}_i$ нь i дүгээр бүлгийн дундаж, $\overline{x}$ нь нийт түүврийн дундаж юм.

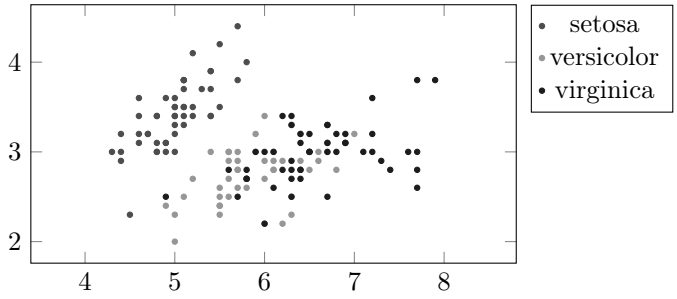
Шинжүүрийн статистикийн асимптот тархалт

$$-\left(n - 1 - \frac{p+k}{2}\right) \ln \Lambda \sim \chi^2_{p(k-1)}$$

Жишээ 28. Цоморлигийн урт болон өргөний дундаж хэмжээ тус гурван зүйлийн хувьд ялгаатай болохыг батлах зорилго тавъя.

Өгөгдөл бэлдэх

```
X <- as.matrix(iris[c("Sepal.Length", "Sepal.Width")])
group <- iris$Species
```



Зураг 11: Цахилдаг цэцгийн setosa, versicolor болон virginica зүйлүүдийн цоморлигийн урт болон өргөн

Олон хэмжээст дисперсийн шинжилгээ

```
summary(manova(X ~ group), test = "Wilks")
```

Үр дүн

```
           Df   Wilks approx F num Df den Df    Pr(>F)
group       2 0.16654   105.88      4    292 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Санамж

Аливаа статистик арга техникийг хэрэглэхдээ урьдач нөхцөл нь биелж буй эсэхийг заавал шалгана. Эс бөгөөс буруу статистик дүгнэлт гарч болно. Тухайлбал олон хэмжээст дисперсийн шинжилгээний хувьд өгөгдөл нь дараах нөхцөл хангах ёстой.

1. $\forall i \in \{1, \ldots, k\} : n_i > p$ буюу бүлэг тус бүрийн хэмжээ нь санамсаргүй хувьсагчдын тоо буюу санамсаргүй векторын хэмжээнээс илүү байна.

2. Түүврийн элементүүд хамааралгүй байна. Өөрөөр хэлбэл туршилтууд хамааралгүй байх шаардлагатай. Үүний эсрэг энгийн жишээ бол нэг объект дээр хоёр өөр эгшинд туршилт тавих явдал юм. Тус шинжилгээний хувьд ийм давтагдсан туршилт байж болохгүй.

3. Онцгой утгагүй байна. Олон хэмжээст дисперсийн шинжилгээ онцгой утгад мэдрэг гэж судлаач нар тэмдэглэсэн нь бий.

4. Өгөгдөл олон хэмжээст хэвийн тархалттай байна.

5. Бүлгүүдийн ковариацийн матриц адил байна.

Сүүлийн гурван нөхцлийг шалгах шинжүүрүүдийг үүний өмнө үзсэн.

Лекц VII
Гол хэсгийн шинжилгээ

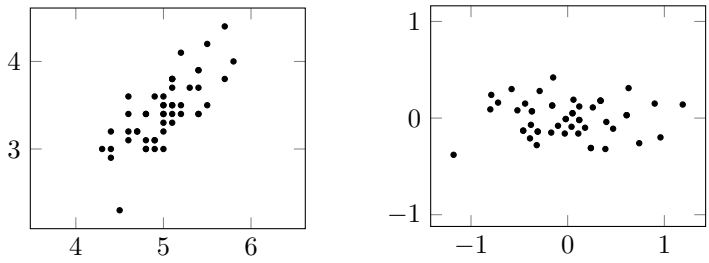
1 Танилцуулга

Гол хэсгийн шинжилгээ

Гол хэсгийн шинжилгээ буюу Principal Component Analysis (PCA) нь санамсаргүй векторын ковариацийн бүтцийг задлан шинжлэх статистикийн арга техник юм. Тус шинжилгээ нь олон жилийн өмнө гарсан сонгодог арга боловч орчин үеийн машин сургалт, өгөгдлийн уурхай, их өгөгдөл боловсруулах зэрэгт хэмжээс бууруулах, гол чухал хүчин зүйлсийг илрүүлэх зэрэг зорилгод өргөн хэрэглэгдэж байна. Энэ шинжилгээгээр өгөгдлийн хувьсан өөрчлөгдөж буй ортогонал чанартай гол чиглэлүүдийг олж улмаар хувьсагчдыг тэдгээр чиглэлийн дагуу шинээр авч үздэг. Ингэснээр дисперсээр нь эрэмбэлсэн хамааралгүй хувьсагчид үүсдэг. Ингээд бага дисперстэй буюу ач холбогдол багатай хувьсагчдыг орхивол энэ нь хэмжээс бууруулсан явдал болно. Өөрөөр хэлбэл дисперс ихтэйг голлох хүчин зүйл гэж чухалчилж үзнэ.

Их дисперстэй чиглэл ба эргүүлэлт

R програмын `datasets` багцад байдаг `iris` датафрейм дахь цахилдаг цэцгийн Setosa зүйлийн цоморлигийн урт болон өргөн гэсэн хоёр хувьсагч дээрх гол хэсгийн шинжилгээний үр дүнг цэгэн диаграммаар дүрсэлж үзүүлэв.



шинжилгээний өмнө шинжилгээний дараа

Зураг 12: Хоёр хэмжээст өгөгдөл дээрх гол хэсгийн шинжилгээ

Эргүүлэлт буюу проекц

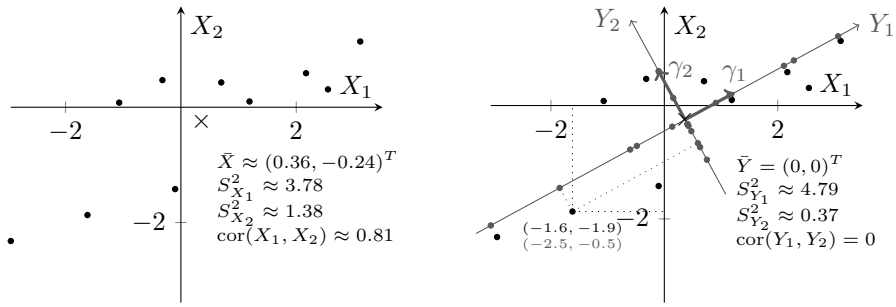
Дисперс

```
X4 <- datasets::iris3[,, "Setosa"]
```

Өгөгдөл

```
head(X4)
```

Үр дүн



(a) Өгөгдөл (b) Үр дүн

Зураг 13: Гол хэсгийн шинжилгээ

```
  Sepal L. Sepal W. Petal L. Petal W.
       5.1      3.5      1.4      0.2
       4.9      3.0      1.4      0.2
       4.7      3.2      1.3      0.2
       4.6      3.1      1.5      0.2
       5.0      3.6      1.4      0.2
       5.4      3.9      1.7      0.4
```

Дисперс

```
apply(X = X4, MARGIN = 2, FUN = var)
```

Үр дүн

```
   Sepal L.   Sepal W.   Petal L.   Petal W.
0.12424898 0.14368980 0.03015918 0.01110612
```

Шинжилгээ

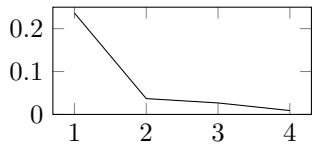
```
pca <- prcomp(X4)
```

Дисперс

```
pca$sdev^2
```

Үр дүн

```
0.236455690 0.036918732 0.026796399 0.009033261
```



2 Гол хэсгийн шинжилгээ

Хамгийн их дисперстэй чиглэл дэх проекц буюу шинэ хувьсагч

Эхлээд $E(X) = 0$ гэж үзнэ. Зохих үр дүнд хүрсэний дараа $E(X) = \mu \neq 0$ тохиолдлыг тусган оруулна.

X санамсаргүй векторын хамгийн ихээр хувьсан өөрчлөгдөж буй чиглэлийг хэрхэн олох вэ? Өөрөөр хэлбэл аль чиглэлд проекцлох вэ?

$$Y = \delta^T X = \delta_1 X_1 + \ldots + \delta_p X_p$$

проекцын $\delta = (\delta_1, \ldots, \delta_p)^T$ вектор ямар байх вэ?

δ нь зөвхөн чиглэл заах тул $\|\delta\|^2 = \delta^T\delta = \sum_{i=1}^{p} \delta_i^2 = 1$ буюу уртыг нь нэгтэй тэнцүү гэнэ.

Ийнхүү дээрх проекцоор үүсэх Y хувьсагчийн дисперсийг хамгийн их

$$D(Y) = D(\delta^T X) \longrightarrow \max$$

байлгах проекцын δ векторыг олъё.

$D(\delta^T X) \longrightarrow \max$ **бодлого**

$$D(\delta^T X) = \operatorname{cov}(\delta^T X) = \delta^T \operatorname{cov}(X)\delta = \delta^T \Sigma \delta$$

Бодлого 1. $\delta^T\delta = 1$ нөхцөлд

$$\max_{\delta} \delta^T \Sigma \delta = ?$$

максимум олох

$D(\delta^T X) \longrightarrow \max$ **бодлогын бодолт**

- Лагранжийн функц ба Лагранжийн нөхцөл

$$\mathrm{L}(\delta, \lambda) = \delta^T \Sigma \delta - \lambda(\delta^T\delta - 1)$$

$$\frac{\partial \mathrm{L}}{\partial \delta} = 2\Sigma\delta - 2\lambda\delta = 0$$

- Хувийн утга

$$\Sigma\delta = \lambda\delta \Longleftrightarrow \delta^T\Sigma\delta = \lambda$$

- Шийд

$$\max_{\delta} \delta^T \Sigma \delta = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Энд λ_i нь Σ ковариацийн матрицын хувийн утга юм. Улмаар $\|\delta\|^2 = 1$ болон $\max_{\delta} \delta^T\Sigma\delta = \lambda_1$ бас $\Sigma = \Gamma\Lambda\Gamma^T$ буюу матрицын хувийн утгын задаргаа зэргийг тооцвол

$$\delta = \gamma_1$$

болно.

Бодолтын үр дүн буюу нэг дүгээр гол хэсэг

- **Хамгийн их диспер бүхий чиглэл**

$$\gamma_1$$

- **Шинэ хувьсагч**

$$Y_1 = \gamma_1^T(X - \mu)$$

Энд $\mu = E(X)$ гэв.

- **Шинэ хувьсагчийн дисперс**

$$D(Y_1) = \lambda_1$$

Дараагийн гол хэсгүүд

Нэмэлт нөхцөл
Дараагийн гол хэсэг өмнө олсон γ_1 чиглэлтэй ортогонал буюу $\delta^T\gamma_1 = 0$ байна.

Бодлого: Дараагийн гол хэсэг
$\delta^T\delta = 1$ ба $\delta^T\gamma_1 = 0$ нөхцөлд $\max_{\delta} \delta^T\Sigma\delta = ?$

$$\mathrm{L}(\delta, \alpha, \beta) = \delta^T\Sigma\delta - \alpha(\delta^T\delta - 1) - \beta(\delta^T\gamma_1)$$

$$\frac{\partial \mathrm{L}}{\partial \delta} = 2\Sigma\delta - 2\alpha\delta - \beta\gamma_1 = 0$$

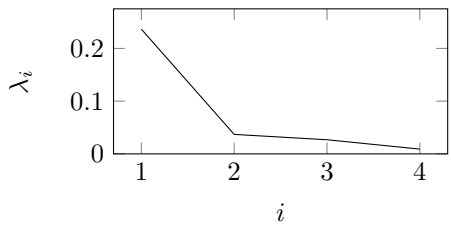
$$\delta^T\Sigma\delta = \alpha$$

Энэ нь өмнөхтэй л адил хувийн утга, хувийн вектор руу хөтлөх бөгөөд λ_1 хувийн утгыг нэгэнт ашигласан тул одоо $\alpha = \lambda_2$ гэнэ.

Гол хэсгүүд, тэдгээрийн дисперс, сайр чулууны диаграмм ба сонгон авах гол хэсгийн тоо

$$Y_i = \gamma_i^T(X - \mu), \qquad \mathrm{cov}(Y_i) = \lambda_i, \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

Гол хэсгийн тоог сонгоход ашигладаг аргачлалуудын нэг нь сайр чулууны



Зураг 14: Хувийн утгуудаар байгуулсан сайр чулууны диаграмм

диаграммд тулгуурладаг бөгөөд гол хэсгүүдийг "голын ёроол" эхэлтэл авдаг.

Бодолтын нэгдсэн үр дүн буюу эцсийн хариу
$E(X) = \mu$ ба $\mathrm{cov}(X) = \Sigma = \Gamma\Lambda\Gamma^T$ бол

$$Y = \Gamma^T(X - \mu)$$

$$\mathrm{cov}(Y) = \Lambda$$

болно.

R програм дээрх гол хэсгийн шинжилгээний функцүүд

R програмын `stats` багцад гол хэсгийн шинжилгээнд зориулсан `prcomp()` болон `princomp()` гэсэн хоёр функц байдаг. Эдгээрээс `prcomp()` функц нь уг сэдвээр танилцуулсан онолтой нийцдэг. Гэхдээ проекцын чиглэл сонголт буюу эргүүлэлт нь тохиолдлынх байдаг тул энэ нь бусад програм бас R програмын өөр хувилбараас гарсан үр дүнгээс тэмдгээрээ ялгаатай байж болно. Харин `princomp()` нь түүврийн дундаж квадрат хазайлт болон проекцын чиглэл олох байдлаараа ялгаатай бөгөөд үр дүн нь S-Plus програмынхтай тохирдог. `prcomp()` нь дараах бүтэцтэй утга буцаана.

sdev гол хэсгүүдийн стандарт хазайлт буюу хувийн утгуудын язгуур - $\lambda_1^{1/2}, \ldots, \lambda_p^{1/2}$

rotation гол хэсгүүд дээрх хувьсагчдын жингүүд буюу хувийн векторуудаас тогтох эргүүлэлтийн матриц - Γ

center тархалтын төв буюу хувьсагчдын дундаж утга - μ

scale масштабын нөлөөг зайлуулсан буюу хувьсагчдын стандарт хазайлтыг нэгтэй тэнцүү болгосон эсэх

x түүврийн элементүүдийн гол хэсгүүд дээрх проекц - Y

`FactoMineR` багцын `PCA()` функц

Гол хэсгийн шинжилгээг бас R програмын `FactoMineR` нэмэлт багцын `PCA()` функцийн тусламжтай хийж болдог. Тус функц `princomp()` функцтэй ижил үр дүн гаргадаг боловч гол хэсэг дээрх хувьсагчдын оролцоо зэрэг бусад нэмэлт утга буцаадаг.

```
FactoMineR::PCA(X, scale.unit = TRUE, ncp = 5, graph = TRUE)
```

Ашигласан аргументуудын тайлбар

X өгөгдөл агуулсан матриц эсвэл датафрейм

scale.unit масштабын нөлөөг зайлуулах буюу хувьсагчдын стандарт хазайлтыг нэгтэй тэнцүү болгож хувиргах эсэх

ncp функцийн буцаах утга дотор оруулах гол хэсгийн тоо

graph түүврийн элементүүд болон хувьсагчдын эхний хоёр гол хэсэг дээрх проекцыг харуулсан диаграмм байгуулах эсэх

`FactoMineR::PCA()` функц өмнө авч үзсэн функцүүдээс илүү дэлгэрэнгүй үр дүн бүхий утга буцаадаг. Тухайлбал гол хэсгүүдийн дисперс буюу түүврийн ковариацийн матрицын хувийн утгуудаас гадна гол хэсгүүдийн ганцаарчилсан болон хуримтлуулах байдлаар бодсон дисперсийн хувиудыг нийлүүлэн матриц болгосон байдаг. Тэрхүү матриц нь тус функцийн буцаах утгын `eig` элементэд агуулагддаг.

Сайр чулууны диаграмм байгуулахад `factoextra` багцын `fviz_eig()` функц ашиглана. Ингэхдээ тус функцийн `X` аргументаар `FactoMineR::PCA()` функцийн утгыг дамжуулна.

Харин Y буюу гол чиглэл дээрх түүврийн элементүүдийн проекц нь `FactoMineR::PCA()` функцийн хувьд түүний утгын `ind` элемент дэх `coord` матрицад агуулагддаг.

`FactoMineR::PCA()` функцийн бусад үр дүнг зохих слайд дээр танилцуулна.

Гол хэсгүүдийн шинж чанар

Чанар 9. 1. $E(Y_i) = 0$

2. $D(Y_i) = \lambda_i$
3. $\operatorname{cov}(Y_i, Y_j) = 0,\ i \neq j$
4. $D(Y_1) \geq \ldots \geq D(Y_p) \geq 0$

Гол хэсгүүдийн шинж чанар ба өргөтгөсөн дисперс

$|\Sigma| = |\operatorname{cov}(X)|$ буюу X санамсаргүй векторын ковариацийн матрицын тодорхойлогчийг X санамсаргүй векторын *өргөтгөсөн дисперс* гэдэг.

Чанар 10. $\prod_{i=1}^{p} D(Y_i) = |\Sigma|$

Баталгаа $|\Sigma| = |\Gamma^T \Lambda \Gamma| = |\Lambda \Gamma \Gamma^T| = |\Lambda| = \prod_{i=1}^{p} \lambda_i = \prod_{i=1}^{p} \operatorname{cov}(Y_i)$ □

$$\underbrace{|\Sigma|}_{\substack{\text{нийт}\\ \text{дисперс}}} = \underbrace{D(Y_1 = \delta^T X)}_{\substack{\text{голлох}\\ \text{дисперс}}} \cdot \underbrace{D(U)}_{\substack{\text{үлдэгдэл}\\ \text{дисперс}}}$$

Гол хэсгүүдийн шинж чанар ба сонгон авах гол хэсгийн тоо

Чанар 11. $\sum_{i=1}^{p} D(Y_i) = \operatorname{tr}(\Sigma)$

Гол хэсгийн тоог сонгох аргачлалуудын нэг нь дээрх чанар дахь гол хэсгүүдийн дисперсийн нийлбэрийн тусламжтай томьёологддог. Ийнхүү q ширхэг гол хэсгээр тайлбарлагдах дисперсийн нийлбэр дисперсэд эзлэх хувь буюу

$$\psi_q = \frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_p} \cdot 100\%$$

харьцааны утга хангалттай их бол гол хэсгийн тоог q гэж сонгоно.

$\lambda_1, \ldots, \lambda_p$ хувийн утгууд болон ψ_q хувиуд `FactoMineR::PCA()` функцийн утгын `eig` элементэд агуулагдана.

Масштабын нөлөө, сонгон авах гол хэсгийн тоог олох Кайзерын дүрэм

Хувьсагчдын хэмжүүрийн нэгж буюу масштаб ялгаатай үед зарим хувьсагч гол хэсгийн дисперсэд хүчээр давамгайлдаг. Ийм нөлөөг арилгахын тулд ихэвчлэн хувьсагчдын дисперсийг нэгтэй тэнцүү болгож стандартчилдаг. Нөгөө талаас энэ нь хувьсагчдыг "тэгш эрхтэй" болгож буй хэрэг юм. Цаашилбал хэмжээс болон масштабын ялгаа нь тухайлбал сантиметр ба миллиметрийн зөрүүгээр ч тогтохгүй бөгөөд хувьсагчийн утга агуулга, мөн чанартай ч холбогдож болно.

Кайзерын дүрэм нь хувийн утгыг нэгээс бага болтол гол хэсгүүдийг нэмж авахыг зааварладаг. Тус аргачлалыг стандартчилсан хувьсагчид дээрх гол хэсгийн шинжилгээнд ашиглавал зохимжтой. Хувьсагчид стандартчлагдсан бөгөөд хамааралгүй байг. Тэгвэл хувьсагчид бүгд бие даасан фактор болох бөгөөд эдгээрийн хувийн утгууд нэгтэй тэнцүү байна. Харин одоо хөндлөнгийн далд фактор гарч ирэн хувьсагчдад нөлөөлсөнөөр холбоо хамаарал үүссэн гэе. Тэгвэл хувийн утгууд адил байхаа болих буюу гол хэсгүүдийн зарим нь давамгайлж, зарим нь дарангуйлагдана. Иймхүү нэгээс их хувийн утгатай давамгайлагч гол хэсгүүдийг л ялгаж авна.

3 Гол хэсгүүдийг тайлбарлах нь

Гол хэсгүүдийг тайлбарлах нь

$X = (X_1, \ldots, X_p)^T$ вектор дахь хувьсагчид нь практикт "зээлийн хүү", "татварын хувь", "инфляц" гэх мэтчилэн тодорхой утга агуулгатай байдаг. Харин гол хэсгийн шинжилгээгээр олдох шинэ $Y = (Y_1, \ldots, Y_p)^T$ санамаргүй вектор дахь хувьсагчид буюу гол хэсгүүд нь

$$Y_i = \gamma_{1,i}(X_1 - \mu_1) + \ldots + \gamma_{p,i}(X_p - \mu_p)$$

гэж $X_1, \ldots, X_p$ хувьсагчдын холимог байдлаар зохиогдох тул тов тодорхой утга агуулгагүй юм. Иймд эдгээрийг юу гэж үзэх вэ, хэрхэн тайлбарлах вэ гэсэн асуулт аяндаа гарна. Түүнчлэн $Y_1, \ldots, Y_p$ хувьсагчдыг анх өгөгдсөн $X_1, \ldots, X_p$ хувьсагчидтай холбож тайлбарлахаас өөр аргагүй юм. Иймхүү холбож тайлбарлахад ашиглаж болох тоон үзүүлэлт бол корреляцын коэффициент юм.

Анхны хувьсагчид ба гол хэсгүүд хоорондын корреляцын матриц

Анхны хувьсагчид ба гол хэсгүүдийн ковариацын матриц

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY^T) - EXEY^T = E(XY^T) = E(X(\Gamma^T(X-\mu))^T) \\ &= E(X(X-\mu)^T\Gamma) = E(XX^T - X\mu^T)\Gamma = \left(E(XX^T) - EXEX^T\right)\Gamma \\ &= \operatorname{cov}(X)\Gamma = \Sigma\Gamma = \Gamma\Lambda\Gamma^T\Gamma = \Gamma\Lambda\end{aligned}$$

Анхны хувьсагчид ба гол хэсгүүдийн корреляцын матриц

$$\operatorname{cor}(X, Y) = D^{-1/2}\Gamma\Lambda\Lambda^{-1/2} = D^{-1/2}\Gamma\Lambda^{1/2}$$

Энд $D = \operatorname{diag}(D(X_1), \ldots, D(X_p))$ буюу $X_1, \ldots, X_p$ хувьсагчдын дисперсээс тогтох диагонал матриц юм.

Энэхүү корреляц нь `FactoMineR::PCA()` функцийн буцаах утгын `var` элемент доторх `cor` дэд элементэд агуулагддаг.

Чанар 12.

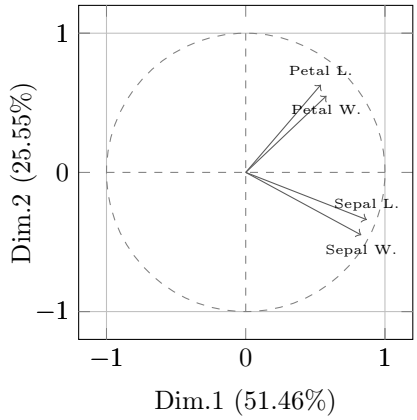
$$[\text{cor}(X_i, Y_1)]^2 + \ldots + [\text{cor}(X_i, Y_p)]^2 = 1$$

Баталгаа $\Sigma = \Gamma^T \Lambda \Gamma$ болохыг анхаарвал

$$[\text{cor}(X_i, Y_1)]^2 + \ldots + [\text{cor}(X_i, Y_p)]^2 = \sum_{j=1}^{p} [\text{cor}(X_i, Y_j)]^2 = \frac{\sum_{j=1}^{p} \lambda_j \gamma_{ij}^2}{\sigma_{X_i X_i}} = \frac{\gamma_i^T \Lambda \gamma_i}{\sigma_{X_i X_i}} = \frac{\sigma_{X_i X_i}}{\sigma_{X_i X_i}} = 1$$

болно. □

Мөрдлөгөө 1. Дурын $q \leq p$ бүрийн хувьд $[\text{cor}(X_i, Y_1)]^2 + \ldots + [\text{cor}(X_i, Y_q)]^2 \leq 1$ байна.



Зураг 15: `X4` өгөгдөл дээр хийсэн шинжилгээнээс олдсон корреляцаар байгуулсан диаграмм

Ийм диаграммыг `plot(x, axes = c(1,2), choix = "varcor")` байдлаар байгуулах бөгөөд `x` аргументаар `FactoMineR::PCA()` функцийн буцаах утгыг дамжуулна.

Гол хэсгүүдийг тайлбарлахтай уялдуулан гол хэсгийн тоог олох

Гол хэсгийн тоо сонгоход хэрэглэж болох өөр нэг арга бол сая үзсэн корреляцын диаграмм болон бусад үр дүнд үндэслэн тайлбарлаж болохуйц утга учиртай үр дүн гартал гол хэсгүүдийг нэмж авах явдал юм.

Түүнчлэн сүүлийн үед параллел шинжилгээ буюу бодит өгөгдлийн хувийн утгуудыг ижил хэмжээтэй санамсаргүй өгөгдлийн хувийн утгуудаас бага болтол гол хэсгүүдийг нэмж авах гэсэн аргачлалыг хүчтэй яригдах болсон. Параллел шинжилгээ нь гол хэсгийн зөв тоог олж тогтоохдоо сайн ч статистикийн програмуудад өргөнөөр тусгагдаагүй л байна.

Гол хэсгийн шинжилгээ ба сингуляр утгын задаргаа

Практикт Σ ковариацийн матрицыг өмнө үзсэнчлэн түүврийн ковариацийн матрицаар үнэлнэ. $E(X) = 0$ нөхцөлд $\mathcal{X}$ түүврийн матрицаар түүврийн ковариацийн матрицыг дараах байдлаар олж болно.

$$S = \frac{1}{n-1}\mathcal{X}^T\mathcal{X}$$

Нөгөө талаас түүрүүн үзсэнчлэн ковариацийн матрицын хувийн утгын задаргааг ашиглавал

$$S = \Gamma\Lambda\Gamma^T = \Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^T = \Gamma\Lambda^{1/2}I\Lambda^{1/2}\Gamma^T = \underbrace{\Gamma\Lambda^{1/2}U^T}_{\frac{1}{\sqrt{n-1}}\mathcal{X}^T}\underbrace{U\Lambda^{1/2}\Gamma^T}_{\frac{1}{\sqrt{n-1}}\mathcal{X}}$$

болно. Иймхүү тус шинжилгээ $\frac{1}{\sqrt{n-1}}\mathcal{X} = U\Lambda^{1/2}\Gamma^T$ сингуляр утгын задаргаатай холбогдоно.

Сингуляр утгын задаргааг бүрдүүлж буй матрицууд `FactoMineR::PCA()` функцийн буцаах утгын `svd` элементийн `U`, `vs`, `V` дэд элементүүдэд агуулагдана.

Гол хэсгүүд дээрх санамсаргүй хувьсагчдын проекц ба оролцоо

Хувьсагчдыг стандартчилсан үед $\frac{1}{\sqrt{n-1}}\mathcal{X} = U\Lambda^{1/2}\Gamma^T$ задаргааны $\Lambda^{1/2}\Gamma^T$ гишүүн буюу $\Gamma\Lambda^{1/2}$ матрицын i дүгээр мөр, j дүгээр баганын элемент нь X_i хувьсагч ба Y_j гол хэсэг хоорондын корреляц коэффициенттэй тэнцэнэ. Түүнчлэн Γ матриц нь эргүүлэлт буюу проекц тодорхойлдог буюу тус матрицын $\gamma_{i,j}$ элемент нь X_i хувьсагчийн Y_j гол хэсэг дээрх скаляр *проекц*ыг заадаг.

Нөгөө талаас $\|\gamma_j\|^2 = \gamma_{1,j}^2 + \ldots + \gamma_{p,j}^2 = 1$ буюу проекцын квадратуудын нийлбэр нэгтэй тэнцүү байдаг. Иймхүү тус нийлбэр дэх $\gamma_{i,j}^2$ нэмэгдэхүүнийг X_i хувьсагчийн Y_j гол хэсэг дээрх *оролцоо* гэдэг.

Гол хэсгүүд дээрх хувьсагчдын проекц ба оролцоо нь `FactoMineR::PCA()` функцийн буцаах утгын `var` элемент доторх `coord` ба `contrib` элементүүдэд агуулагддаг.

Бас $D(Y_j) = \lambda_j$ болохыг анхаарвал хувьсагчдыг стандартчилсан үед

$$\frac{\mathrm{cor}(X_1, Y_j)^2 + \ldots + \mathrm{cor}(X_p, Y_j)^2}{D(Y_j)} = \frac{(\gamma_{1,j}^2 + \ldots + \gamma_{p,j}^2)\lambda_j}{\lambda_j} = 1$$

байх тул хувьсагчдыг стандартчилсан үед X_i хувьсагчийн Y_j гол хэсэг дээрх оролцоог

$$C_{i,j}^{\mathrm{var}} = \frac{\mathrm{cor}(X_i, Y_j)^2}{D(Y_j)} \cdot 100\% = \gamma_{i,j}^2 \cdot 100\%$$

дисперсийн харьцаагаар тодорхойлж болно. Үүний дээрх адилтгал ёсоор тухайн нэг гол хэсгийн хувьд бүх хувьсагчдын оролцооны нийлбэр 100% байна. Өөрөөр хэлбэл

$$\sum_{i=1}^{p} C_{i,j}^{\mathrm{var}} = 100\%, \quad \forall j = 1, \ldots, p$$

байна.

Гол хэсгүүд дээрх түүврийн элементүүдийн проекц ба оролцоо

Түүврийн нэг элемент буюу үүнд харгалзах $\mathcal{X}$ матрицын тухайн нэг мөр дээр гол хэсгийн шинжилгээний $Y = \Gamma^T(X - \mu)$ хувиргалт хийхэд гарах үр дүн нь уг түүврийн элементийн *проекц* юм.

Түүврийн i дүгээр элементийн j дүгээр гол хэсэг дээрх *оролцоо*г тус гол хэсгийн дисперс буюу дундаж квадрат хазайлт дахь тухайн элементийн хазайлтын квадратын эзлэх хувиар тодорхойлдог. $E(Y_j) = 0$ буюу гол хэсгийн дундаж тэгтэй тэнцүү бас $D(Y_j) = \lambda_j$ ба $\widehat{D(Y_j)} = S^2(Y_j)$ буюу эх олонлогийн дисперсийг түүврийн дисперсээр үнэлдэг зэргийг тооцвол уг оролцоог дараах байдлаар томьёолж болно.

$$C_{i,j}^{\text{ind}} = \frac{\frac{1}{n-1}y_{i,j}^2}{\frac{1}{n-1}\sum\limits_{i=1}^{n} y_{i,j}^2} \cdot 100\% = \frac{\frac{1}{n-1}y_{i,j}^2}{S^2(Y_j)} \cdot 100\% = \frac{\frac{1}{n-1}y_{i,j}^2}{\lambda_j} \cdot 100\%$$

Түүврийн элементүүдийн гол хэсгүүд дэх оролцоо `FactoMineR::PCA()` функцийн буцаах утгын ind элементийн `contrib` дэд элементэд агуулагдана.

Санамсаргүй хувьсагчид болон түүврийн элементүүдийн гол хэсгүүд дээрх проекцын диаграмм байгуулах нь

Хувьсагчдын гол хэсгүүд дээрх проекцын диаграммыг `plot(x, axes = c(1,2), choix = "var")` байдлаар байгуулах бөгөөд `x` аргументаар `FactoMineR::PCA()` функцийн буцаах утгыг дамжуулна. Харин түүврийн элементүүдийн гол хэсэг дээрх проекц буюу Y хувьсагчдаар цэгэн диаграмм байгуулах бол `choix` аргументаар `"ind"` утга дамжуулсан тушаал өгнө.

Лекц VIII
Факторын шинжилгээ I

1 Факторын шинжилгээний загвар

Факторын шинжилгээний тухай ерөнхий ойлголт

Факторын шинжилгээ буюу Factor Analysis нь өмнө үзсэн гол хэсгийн шинжилгээ(Principal Component Analysis)-тэй хэмжээс бууруулах, бие даасан нуугдмал хүчин зүйлсийг олж илрүүлэх зэргээрээ төстэй боловч "гол хэсэг"-ийн тоог нь урьдаас зааж өгдгөөрөө ялгаатай. Мөн тус хоёр шинжилгээ хоёулаа хувьсагчдын холбоо хамаарлыг задалж тайлбарлах зорилготой боловч загварын томьёоллын хувьд эрс ялгаатай. Факторын шинжилгээг хэрэглэх зорилгоос нь хамааруулж *хайгуулын* болон *нотолгооны* гэж хоёр ангилдаг. Хайгуулын факторын шинжилгээнд факторын тоо хатуу бэхлэгдээгүй байх бөгөөд судлаач факторын тоо болон бусад зүйлсээс хамаарч тодорхойлогдох янз бүрийн загваруудыг харьцуулж хамгийн оновчтойг нь сонгох зорилго тавина. Харин нотолгооны факторын шинжилгээг сонгон авсан факторын тоо болон бусад таамаг төсөөллөө батлах эсвэл няцаах зорилгоор хийнэ.

Факторын шинжилгээний загвар

μ дундаж утгын вектор, Σ ковариацийн матриц бүхий

$$X = (X_1, \ldots, X_p)^T$$

санамсаргүй вектор авч үзье.

Загвар

$$X = LF + \mu \quad EF = 0 \quad \operatorname{cov}(F) = I_k \quad k \le p$$

$$F = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix} \quad L = \begin{pmatrix} \\ q_{ij} \\ \\ \end{pmatrix}_{i=\overline{1,p};\, j=\overline{1,k}}$$

$X = LF + \mu$ нь X_i хувьсагчийн хувьд $X_i = q_{i1}f_1 + \ldots + q_{ik}f_k + \mu_i$ байна.

Факторын ачилт гэх нэр томьёоны тухай

L буюу факторын шинжилгээний загвар дахь коэффициентуудыг *факторын ачилт* гэдэг. $X = LF + \mu$ загварыг X векторын ямар нэг X_i компонентийн хувьд бичвэл

$$X_i = q_{i1}f_1 + \ldots + q_{ik}f_k + \mu_i$$

байна. Энд q_{ij} нь X_i хувьсагчийн f_j фактор дээрх "жин" буюу X_i хувьсагчид агуулагдах мэдээллээс f_j факторт хуваарилагдан ачигдаж буй хэсэг юм.

Ковариацийн матриц ба факторын ачилт

$$\begin{aligned} \Sigma &= E(X-\mu)(X-\mu)^T \\ &= E(LF(LF)^T) \\ &= LE(FF^T)L^T \\ &= L\operatorname{cov}(F)L^T \\ &= LL^T \end{aligned}$$

Факторуудыг тайлбарлах нь

Ковариацийн матриц

$$\begin{aligned} \operatorname{cov}(X,F) &= E\{(X-\mu)(F-EF)^T\} \\ &= E\{LFF^T + UF^T\} \\ &= LE\{FF^T\} + E\{UF^T\} \\ &= L\operatorname{cov}(F) \\ &= L \end{aligned}$$

Корреляцийн матриц

$$\operatorname{cor}(X,F) = D^{-1/2}L$$

энд $D = \operatorname{diag}(\sigma_{X_1X_1}, \ldots, \sigma_{X_pX_p})$

2 Шугаман хамааралтай хувьсагчид дээрх шинжилгээ

Шугаман хамааралтай хувьсагчид дээрх шинжилгээ

$$X = (\underbrace{X_1, \ldots, X_k}_{\text{шугаман хамааралгүй}} , \underbrace{X_{k+1}, \ldots, X_p}_{X_1,\ldots,X_k \text{ хувьсагчдаас хамааралтай}})^T$$

Тэгвэл Σ матрицын хувийн утгын задаргаа

$$\Sigma = \begin{pmatrix}\Gamma_1 & \Gamma_2\end{pmatrix} \begin{pmatrix}\Lambda_1 & 0 \\ 0 & 0\end{pmatrix} \begin{pmatrix}\Gamma_1 & \Gamma_2\end{pmatrix}^T$$

байх бөгөөд энд

$$\Lambda_1 = \begin{pmatrix}\lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_k\end{pmatrix}$$

$\lambda_{k+1} = \ldots = \lambda_p = 0$ ба Γ_2 нь эдгээрт харгалзах хувийн векторуудаас тогтох матриц юм.

Сэргээн санах нь 2. Гол хэсгийн шинжилгээ дэх $Y = \Gamma^T(X - \mu)$ хувиргалт

$$Y = \begin{pmatrix}Y_1 \\ Y_2\end{pmatrix} = \begin{pmatrix}\Gamma_1 & \Gamma_2\end{pmatrix}^T (X - \mu)$$

$EY = 0$, $\text{cov}(Y) = \begin{pmatrix}\Lambda_1 & 0 \\ 0 & 0\end{pmatrix}$ тул $EY_2 = 0$ ба $\text{cov}(Y_2) = 0$ болно. Иймд

$$P(Y_2 = 0) = 1$$

дүгнэлт гарна. Улмаар дээрх бүгдийг тооцвол

$$X - \mu = \Gamma Y = \begin{pmatrix}\Gamma_1 & \Gamma_2\end{pmatrix} \begin{pmatrix}Y_1 \\ Y_2\end{pmatrix} = \Gamma_1 Y_1 + \Gamma_2 Y_2 = \Gamma_1 Y_1$$

болно. Ийнхүү

$$X = \Gamma_1 Y_1 + \mu$$

үр дүнд хүрнэ.

Шийд

$$X = \underbrace{\Gamma_1 \Lambda_1^{1/2}}_{L} \underbrace{\Lambda_1^{-1/2} Y_1}_{F} + \mu$$

Загварын нөхцөл хангах эсэхийг шалгах

$$\begin{aligned} EF &= E(\Lambda_1^{-1/2} Y_1) \\ &= \Lambda_1^{-1/2} E Y_1 \\ &= \Lambda_1^{-1/2} E(X_1 - \mu_1) \\ &= 0 \end{aligned}$$

$$\begin{aligned}\operatorname{cov}(F) &= \operatorname{cov}(\Lambda_1^{-1/2} Y_1) \\ &= \Lambda_1^{-1/2} \operatorname{cov}(Y_1)(\Lambda_1^{-1/2})^T \\ &= \Lambda_1^{-1/2} \operatorname{cov}(Y_1)\Lambda_1^{-1/2} \\ &= \Lambda_1^{-1/2}\Lambda_1\Lambda_1^{-1/2} \\ &= I_k\end{aligned}$$

3 Шугаман хамааралгүй хувьсагчид дээрх шинжилгээ

Шугаман хамааралгүй хувьсагчид дээрх шинжилгээ

Энэ тохиолдолд шууд орхиж болох хувьсагч байхгүй тул X санамсаргүй векторын ковариацийн зарим хэсгийг орхихоос өөр аргагүй юм.

Загвар

$$\begin{array}{ll} X = LF + U + \mu & \operatorname{cov}(F, U) = 0 \\ EU = 0 & \operatorname{cov}(U) = \operatorname{diag}(\psi_1, \ldots, \psi_p) \end{array}$$

Энд U нь "хөндлөнгийн" санамсаргүй хүчин зүйлийн нөлөөг илэрхийлнэ.

Дээрх загвар X_i тухайн нэг хувьсагчийн хувьд

$$X_i = q_{i1} f_1 + \ldots + q_{ik} f_k + U_i + \mu_i, \quad i = 1, \ldots, p$$

хэлбэрээр бичигдэнэ.

Ковариацийн задаргаа

Нийт ковариацийн задаргаа

$$\underbrace{\Sigma}_{\text{нийт ковариац}} = \underbrace{LL^T}_{\text{тайлбарлагдах ковариац}} + \underbrace{\Psi}_{\text{орхигдох ковариац}}$$

X_i хувьсагчийн дисперс буюу дундаж квадрат хазайлтын задаргаа

$$\operatorname{cov}(X_i) = \underbrace{q_{i1}^2 + \ldots + q_{ik}^2}_{\| \; h_i^2} + \psi_i$$

Энд h_i^2 нь X_i хувьсагчийн факторуудаар тайлбарлагдах дисперсийн хэмжээг илтгэх бөгөөд үүнийг *нийлэмж* гэдэг. LL^T матрицын гол диагонал дээр h_i^2 $(i = 1, \ldots, p)$ нийлэмж байрлана.

Загварыг үнэлж олох буюу факторын шинжилгээний ажиллах зарчим

Тус шинжилгээ нь факторуудын тоо ширхэг буюу хэмжээсд тохируулж, $\Sigma = LL^T + \Psi$ задаргаа дээр тулгуурлаж

- эсвэл Ψ буюу орхих
- эсвэл L буюу ялган авч үлдээх

ковариацаа эхэлж олоод улмаар нөгөө ковариацаа олох байдлаар явагддаг. Загварын үнэлгээний арга техникүүдийг дараагийн хичээлээр үзнэ.

4 Загварын онцлог

Факторын шинжилгээний загвар масштабаас үл шалтгаалах нь

X санамсаргүй векторын масштаб өөрчлөх нь $C = \mathrm{diag}(c_1, \ldots, c_p)$ диагонал матрицаар $Y = CX$ гэж хувиргахтай адил юм. Энэ тохиолдолд

$$\mathrm{cov}(Y) = C\Sigma C^T = C(L_X L_X^T + \Psi_X)C^T = \underbrace{CL_X}_{L_Y} L_X^T C^T + \underbrace{C\Psi_X C^T}_{\Psi_Y}$$

буюу Y хувьсагчийн хувьд факторын шинжилгээний загвар мөн адил хүчинтэй. Иймд X векторыг $Y = D^{-1/2}(X - \mu)$ стандарт хувиргалтаар хувиргалаа ч загвар хүчинтэй байна. Энэ тохиолдолд $\mathrm{cov}(Y) = \mathrm{cor}(Y)$ буюу ковариац болон корреляцийн матрицууд адил байх тул факторын шинжилгээнд корреляцын матрицыг түлхүү ашигладаг.

Факторын ачилтын цор ганц бус байдал ба эргүүлэлт

Хэрэв J нь ортогонал буюу $JJ^T = I$ чанартай матриц бол загварыг дараах байдлаар өөрчлөн бичиж болно.

$$X = LF + U + \mu = \underbrace{(LJ)}_{\text{ачилт}} \underbrace{(J^T F)}_{\text{фактор}} + U + \mu$$

F векторыг ортогонал матрицаар үржүүлэх нь координатын тэнхлэгийг эргүүлэхтэй ижил юм. $\mathrm{cor}(X, F) = D^{-1/2}L$ байсан тул уг эргүүлэлтийг тус корреляцыг хамгийн их байлгах гэх мэтчилэн ашигтайгаар сонговол зохимжтой. Ийнхүү эргүүлэлтийн тодорхойгүй байдал буюу факторын ачилтын олон утгат байдлаас зайлсхийхийн тулд

$$L^T\Psi^{-1}L \text{ диагонал} \qquad \text{буюу} \qquad L^T D^{-1} L \text{ диагонал}$$

гэсэн нэмэлт нөхцөл тавьдаг. Энэ нь $D^{-1/2}L$ буюу хоёр өөр факторын стандартчилсан ачилт ортогонал гэсэн санааг илэрхийлнэ. Иймд хэрэв шинжилгээнд корреляцийн матриц ашигласан буюу хувьсагчдыг стандарт хувиргалтаар хувиргасан гэвэл $D = I_p$ болох тул факторуудын ачилт L нь шууд ортогонал чанартай болно.

$\Sigma = LL^T + \Psi$ загвар шийдтэй эсэхийг тогтоохын тулд чөлөөний зэргийг нь шинжилнэ.

$$d = (\text{нөхцөлгүй үеийн параметрүүдийн тоо})$$

$$
\begin{aligned}
&\quad - \text{ (нөхцөлтэй үеийн параметрүүдийн тоо)} \\
&= (\Sigma \text{ матрицын элементүүдийн тоо}) \\
&\quad - \text{ (загварын параметрүүдийн тоо)} \\
&= \frac{1}{2}p(p+1) - \left(pk + p - \frac{1}{2}k(k-1)\right) \\
&= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)
\end{aligned}
$$

$d < 0$ загвар тодорхойлогдохгүй

$d = 0$ эргүүлэлтгүйгээр бүрэн шийдэгдэнэ

$d > 0$ яг таг шийд оршин байхгүй. Энэ тохиолдолд $\Sigma = LL^T + \Psi$ загварт тулгуурласан ойролцоо бодолт хийдэг.

Жишээ 29. $p = 3, 4, 6$ үед $\forall k \leq p$ бүрт харгалзах чөлөөний зэрийг олж зохих дүгнэлт гарга.

$p = 3$ үед

$$d = \begin{cases} 0, & k = 1 \\ \leq -2, & k \geq 2 \end{cases}$$

$p = 4$ үед

$$d = \begin{cases} 2, & k = 1 \\ \leq -1, & k \geq 2 \end{cases}$$

$p = 6$ үед

$$d = \begin{cases} 9, & k = 1 \\ 4, & k = 2 \\ 0, & k = 3 \\ \leq -3, & k \geq 4 \end{cases}$$

5 Факторын тоо

Факторын тоо

Факторын зохимжит тоог тогтоох нь факторын шинжилгээний гол асуудлуудын нэг юм. Ийм олон аргачлал байдаг ч төгс арга үгүй л байна.

Факторын тоо тогтоох аргачлалуудаас

1. Сайр чулууны шинжүүр буюу хувийн утгуудаар байгуулсан диаграмм дээрх уналтын налуугаас голын эрэг ба сайр эсвэл уулын налуу ба бэл мэт хэлбэрийг ажиглах

2. Хувийн утгыг 1-ээс бага болтол гол хэсгүүдийг задлах

3. Тайлбарлаж болохуйц утга учиртай үр дүн гартал факторуудыг задлах

4. Параллел шинжилгээ буюу бодит өгөгдлийн хувийн утгуудыг ижил хэмжээтэй санамсаргүй өгөгдлийн хувийн утгуудаас бага болтол факторуудыг задлах

Лекц IX
Факторын шинжилгээ II

1 Загварын үнэлгээ

Загварын үнэлгээ

Факторын шинжилгээний

$$X = LF + U + \mu$$

загвар дахь L болон F үл мэдэгдэх параметрүүдийг хэрхэн үнэлэхийг авч үзье. Үнэлгээг эхлээд дээрх загвараас мөрдөн гарах

$$\Sigma = LL^T + \Psi$$

задаргаанаас L ачилтыг олох улмаар F факторыг олох гэсэн дарааллаар хийнэ.

Сэргээн санах нь 3. Факторын шинжилгээний $X = LF + U + \mu$ загвар ашиглаж гаргасан ковариацийн задаргааг шинжилгээний загвар гэж ойлгож болох бөгөөд нэг дүгээрт

$$\Sigma = LL^T + \Psi$$

задаргаа, хоёр дугаарт тус задаргаа дахь L ба Ψ үл мэдэгдэх параметрүүдтэй холбогдох

$$L^T\Psi^{-1}L \text{ диагонал} \qquad \text{буюу} \qquad L^TD^{-1}L \text{ диагонал}$$

нэмэлт нөхцөл хоёрт байх мэдэгдэгч ба үл мэдэгдэгчдийн тоо ширхэгийн зөрүүг чөлөөний зэрэг гэнэ.

- $d < 0$ үед загвар тодорхойлогдохгүй.
- $d = 0$ үед цор ганц утгатай шийд олдоно.
- $d > 0$ үед ковариацийн задаргаанд тулгуурласан ойролцоо бодолт хийж шийд олно.

Сэргээн санах нь 4. Факторын ачилт L нь масштабаас үл шалтгаална.

Дээрх чанараас үүдэн загвар үнэлэхэд ковариацийн матриц ба корреляцийн матрицын алийг нь ч ашиглаж болно. Ковариацийн матриц ашиглах үед

$$S = \hat{L}\hat{L}^T + \hat{\Psi}$$

тэгшитгэл, корреляцын матриц ашиглах үед

$$R = \hat{L}\hat{L}^T + \hat{\Psi}$$

тэгшитгэл ашиглана. Энд S болон R нь харгалзан түүврийн ковариацийн матриц болон түүврийн корреляцын матриц юм.

Жишээ 30. $p = 3$ ба $k = 1$ тохиолдолд загварын үнэлгээ хий.

$d = \frac{1}{2}(3-1)^2 - \frac{1}{2}(3+1) = 0$ тул загвар цор ганц шийдтэй. Корреляцын матриц ашиглавал

$$\begin{array}{ccc} R & = & \hat{L}\hat{L}^T + \hat{\Psi} \\ \| & & \| \\ \begin{pmatrix} 1 & r_{X_1X_2} & r_{X_1X_3} \\ r_{X_1X_2} & 1 & r_{X_2X_3} \\ r_{X_1X_3} & r_{X_2X_3} & 1 \end{pmatrix} & & \begin{pmatrix} \hat{q}_1^2 + \hat{\psi}_{11} & \hat{q}_1\hat{q}_2 & \hat{q}_1\hat{q}_3 \\ \hat{q}_1\hat{q}_2 & \hat{q}_2^2 + \hat{\psi}_{22} & \hat{q}_2\hat{q}_3 \\ \hat{q}_1\hat{q}_3 & \hat{q}_2\hat{q}_3 & \hat{q}_3^2 + \hat{\psi}_{33} \end{pmatrix} \end{array}$$

үнэлгээний тэгшитгэл бичиж болно. Дээрх тэгшитгэлээс

$$\frac{r_{X_1X_2}r_{X_1X_3}}{r_{X_2X_3}} = \frac{\hat{q}_1\hat{q}_2\hat{q}_1\hat{q}_3}{\hat{q}_2\hat{q}_3} = \hat{q}_1^2$$

уялдаа холбоо ажиглаглана уу. Бас $\hat{q}_2^2$ болон $\hat{q}_3^2$ ачилтуудыг ч дээрх байдлаар олж болно.

U_j хөндлөнгийн хүчин зүйлсийн дисперсүүдийг

$$R = \hat{L}\hat{L}^T + \hat{\Psi}$$

тэгшитгэлийн

$$\begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} = \begin{pmatrix} \hat{q}_1^2 + \hat{\psi}_{11} & & \\ & \hat{q}_2^2 + \hat{\psi}_{22} & \\ & & \hat{q}_3^2 + \hat{\psi}_{33} \end{pmatrix}$$

хэсгээс

$$\hat{\psi}_{jj} = 1 - \hat{q}_j^2$$

байдлаар олно.

Жишээ 31. R програмын `datasets` багц дотор байдаг `mtcars` датафреймын эхний $p = 7$ хувьсагч дээр $k = 2$ ширхэг фактортай загвар тавьж үнэлгээ хий.

Загварын чөлөөний зэрэг

$$\begin{aligned} d &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \\ &= \frac{1}{2}(7-2)^2 - \frac{1}{2}(7+2) \\ &= 8 \end{aligned}$$

буюу эерэг утгатай байгаа тул статистик ач холбогдолтой загвар тодорхойлогдоно.

Өгөгдөл ялгаж авах

```
X <- datasets::mtcars[1:7]
```

Түүврийн хэмжээ

```
nrow(X)
```

```
32
```

Эхний 6 мөрийг нь хэвлэх

```
head(X)
```

```
                   mpg cyl disp  hp drat    wt  qsec
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02
Datsun 710        22.8   4  108  93 3.85 2.320 18.61
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02
Valiant           18.1   6  225 105 2.76 3.460 20.22
```

Загварын үнэлгээний аргуудаас

Гол факторын арга Загварын параметрүүдийг хөндлөнгийн хүчин зүйлсийн дисперсээс эхэлж дараалан дөхөх аргаар олдог. Тус аргын хувьд корреляцийн матриц ашигласан буюу хувьсагчид дээр стандарт хувиргалт хийсэн үед хувьсагчдын дисперсээс тогтох диагонал матриц $D = \text{diag}(s_{X_1,X_1}, \ldots, s_{X_p,X_p}) = I$ болох тул

$$L^T D^{-1} L \text{ диагонал}$$

нэмэлт нөхцөл

$$L^T L \text{ диагонал}$$

хэлбэрт шилжинэ. Иймд L матрицын багануud ортогонал болох бөгөөд улмаар тэдгээрийг $R - \Psi$ матрицын хувийн векторуудаар авах боломжтой болно.

Хамгийн их үнэний хувь бүхий арга X санамсаргүй векторын хамтын тархалт мэдэгдэж байгаа тухайлбал $X \sim N_p(\mu, \Sigma)$ үед хэрэглэнэ. Үүнтэй холбогдуулан k ширхэг фактортай загвар хүчинтэй эсэхийг шалгах үнэний хувийн харьцаат шинжүүр зохиосон байдаг.

Гол хэсгийн арга Гол хэсгийн шинжилгээ дээр үзэж байсан ковариацийн юм уу корреляцийн матрицын хувийн утгын задаргаанаас факторын ачилтыг үнэлдэг.

Гол хэсгийн аргаар загвар үнэлэх

Факторын шинжилгээг корреляцийн матриц дээр тулгуурлаж хийнэ гэе. Жишээний өгөгдөл дэх хувьсагчид ч масштабын хувьд олон янз тул анхнаасаа корреляцийн матриц ашиглавал зохимжтой.

1. **Түүврийн корреляцийн матрицын хувийн утгын задаргаа**

$$R = \Gamma \Lambda \Gamma^T$$

2. **Факторын ачилтын үнэлэлт** Эхний өөрөөр хэлбэл хамгийн их утгатай k ширхэг хувийн утга болон тэдгээрт харгалзах хувийн вектороор дараах үнэлэлт байгуулна.

$$\hat{L} = (\sqrt{\lambda_1}\gamma_1, \ldots, \sqrt{\lambda_k}\gamma_k)$$

3. **Хөндлөнгийн хүчин зүйлсийн дисперсийн үнэлгээ** Ψ нь диагонал матриц тул түүний элемент болох ψ_{jj} дисперсүүдийг $R - \hat{L}\hat{L}^T$ матрицын диагоналын элементүүдээр үнэлнэ.

Гол хэсгийн аргын үндсэн санаа

Эхний k ширхэг хувийн утгуудаар Λ_F, тэдгээрт харгалзах хувийн векторуудаар Γ_F матриц, үлдэх бусдаар нь Λ_U болон Γ_U матрицууд зохиоё.

$$\begin{aligned} R &= \Gamma\Lambda\Gamma^T \\ &= \begin{pmatrix}\Gamma_F & \Gamma_U\end{pmatrix}\begin{pmatrix}\Lambda_F & 0\\ 0 & \Lambda_U\end{pmatrix}\begin{pmatrix}\Gamma_F & \Gamma_U\end{pmatrix}^T = \Gamma_F\Lambda_F\Gamma_F^T + \Gamma_U\Lambda_U\Gamma_U^T \\ &= \underbrace{\Gamma_F\Lambda_F^{1/2}}_{\hat{L}}\underbrace{\Lambda_F^{1/2}\Gamma_F^T}_{\hat{L}^T} + \underbrace{\operatorname{diag}(\Gamma_U\Lambda_U\Gamma_U^T)}_{\hat{\Psi}} + \underbrace{\text{off-diag}(\Gamma_U\Lambda_U\Gamma_U^T)}_{\text{загварын алдаа}} \end{aligned}$$

Иймхүү

$$\begin{aligned} \hat{L} &= \Gamma_F\Lambda_F^{1/2} = (\sqrt{\lambda_1}\gamma_1, \ldots, \sqrt{\lambda_k}\gamma_k) \\ \hat{\Psi} &= \operatorname{diag}(\Gamma_U\Lambda_U\Gamma_U^T) = \operatorname{diag}(R - \hat{L}\hat{L}^T) \end{aligned}$$

үнэлэлт гарна.

Түүврийн корреляцийн матриц

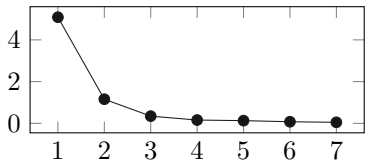
```
R <- cor(X)
```

Хувийн утга болон хувийн вектор

```
eig <- eigen(R)
Lambda <- diag(eig$values)
Gamma <- eig$vectors
```

Хувийн утгууд

```
5.08609988 1.15656554 0.34485150 0.15793358 0.12949405 0.07585706
    0.04919838
```



$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_7} \cdot 100\% \approx 89.18\%$$

Факторын ачилт $\hat{L} = \Gamma_F\Lambda_F^{1/2} = (\sqrt{\lambda_1}\gamma_1, \ldots, \sqrt{\lambda_k}\gamma_k)$

```
L <- Gamma[,1:k] %*% sqrt(Lambda[1:k,1:k])
```

```
           [,1]        [,2]
[1,]  0.9308661 -0.08921934
[2,] -0.9578708 -0.08435906
[3,] -0.9528463  0.08861521
[4,] -0.8744936 -0.36238351
[5,]  0.7468681 -0.48242879
[6,] -0.8825092  0.34667931
[7,]  0.5410937  0.80584767
```

Нийлэмж

Сэргээн санах нь 5. $\hat{h}_j^2 = \sum_{i=1}^{k} \hat{q}_{ji}^2$ нийлэмж нь X_j хувьсагчийн факторуудаар тайлбарлагдах дисперсийн хэмжээ бөгөөд LL^T матрицын гол диагонал дээр байдаг.

```
diag(L %*% t(L))
```

```
0.8744717 0.9246329 0.9157686 0.8960609 0.7905494 0.8990089
    0.9421728
```

Хөндлөнгийн хүчин зүйлсийн дисперсийн үнэлгээ

```
diag(R - L %*% t(L))
```

```
  mpg   cyl  disp    hp  drat    wt  qsec
0.125 0.075 0.084 0.104 0.209 0.101 0.058
```

R програмын `psych` багц дахь `principal()` функц

Шинжилгээ хийх

```
fa <- psych::principal(r = X, nfactors = k, rotate = "none")
```

Шинжилгээний үр дүнгийн тойм

```
print(fa)
```

Үүнээс чөлөөний зэрэг, түүврийн хэмжээ, хувийн утгууд, факторын ачилт, нийлэмж, хөндлөнгийн хүчин зүйлийн дисперс зэргийг дараах байдлаар олно.

```
fa$dof
fa$n.obs
fa$values
```

```
fa$loadings
fa$communality
fa$uniquenesses
```

Гол хэсгийн аргаар үнэлсэн загварын алдаа

R болон $\hat{L}\hat{L}^T + \hat{\Psi}$ матрицуудын гол диагоналын элементүүд тэнцүү, харин диагоналын бус элементүүд үнэлгээнд хагас дутуу оролцож байсан. Тэгвэл загварын алдаа ямар байх вэ?

$$\|R - \hat{L}\hat{L}^T - \hat{\Psi}\|_F = \sqrt{\sum_{i,j} (R - \hat{L}\hat{L}^T - \hat{\Psi})_{ij}^2} \leq \sqrt{\lambda_{k+1}^2 + \ldots + \lambda_p^2} = \|(\lambda_{k+1}, \ldots, \lambda_p)\|_F$$

Алдааны цар хэмжээг олоход Фробениусын норм ашиглав. Эдгээр нормыг дараах байдлаар олж улмаар тэнцүү байгааг нь ажиглаж болно.

```
norm(fa$residual, type = "F")
```

```
norm(as.matrix(fa$values[-{1:k}]), type = "F")
```

Загварын алдааг `psych::principal()` функцийн үр дүнд тусгахын тулд `residuals` аргументаар `TRUE` утга дамжуулна.

2 Факторын үнэлгээ

Факторын үнэлгээ

Факторын үнэлэлтийг *факторын оноо* гэдэг. Өмнөх хэсэгт факторын ачилтыг олохдоо корреляцийн матриц ашигласан. Угтаа энэ нь хувьсагчдын масштабын ялгааг арилгасан явдал юм. Харин факторын оноо олоход факторын ачилт хэрэг болно. Иймд X векторыг L ачилттай нийцүүлэх шаардлагатай. Өөрөөр хэлбэл өгөгдөл дэх хувьсагчдын мастабын ялгааг арилгах хэрэгтэй. Ийнхүү $X_1, \ldots, X_p$ хувьсагч тус бүрийг стандарт хувиргалтаар хувиргасан буюу $EX_i = 0$ ба $DX_i = 1$ гэе. Тэгвэл

$$\operatorname{cov}\begin{pmatrix} X \\ F \end{pmatrix} = \operatorname{cor}\begin{pmatrix} X \\ F \end{pmatrix} = \begin{pmatrix} LL^T + \Psi & L \\ L^T & I_k \end{pmatrix}$$

болно.

X болон F хамтдаа хэвийн тархалттай гэвэл олон хэмжээст хэвийн тархалт сэдэвт үзсэнчлэн F факторыг X санамсаргүй векторын нөхцөл дэх математик дунджаар нь үнэлж болно. $X_1, \ldots, X_p$ хувьсагчдыг стандарт хувиргалтаар хувиргасан тул $\mu = 0$ ба Σ нь корреляцийн матрицтай тэнцүү болно. Иймд тус үнэлэлт $E(F|X = x) = L^T\Sigma^{-1}X$ хэлбэртэй болох тул

$$\hat{F} = L^T R^{-1} X$$

үнэлгээ гарна. Өөрөөр хэлбэл факторыг регрессийн шугаман загварын тусламжтай үнэлнэ.

```
F <- t(t(L) %*% solve(R) %*% t(scale(X)))
```

`psych::principal()` функцийн хувьд факторын оноог түүний буцаах утгын `scores` элемент агуулдаг.

3 Эргүүлэлт

Эргүүлэлт

Факторын ачилтын цор ганц бус байдалтай холбогдуулан эргүүлэлтийн талаар яригдаж байсан билээ.

Сэргээн санах нь 6. $X_1, \ldots, X_p$ хувьсагчдыг стандарт хувиргалтаар хувиргасан буюу шинжилгээнд корреляцийн матриц ашигласан үед $\operatorname{cor}(X, F) = L$ байдаг.

Эргүүлэлтийг $\operatorname{cor}(X, F) = L$ корреляцын абсолют утгыг хамгийн их байлгахаар сонговол ашигтай. Тийм эргүүлэлтийн нэг бол *varimax* эргүүлэлт юм. Тус эргүүлэлт нь корреляцыг зүгээр нэг максимумчлаад зогсолгүйгээр тухайн нэг хувьсагчийг аль болох цөөн факторт ачаалахыг зорьдог. Өөрөөр хэлбэл *varimax* эргүүлэлт нь хувьсагчийг цөөн фактороор тайлбарлах боломж олгодог.

Varimax эргүүлэлт

Varimax эргүүлэлт нь факторуудын ортогонал чанарыг хэвээр хадгалах бөгөөд уг зарчмаар олдох ортогонал эргүүлэлтийг J гэе. Тэгвэл тус эргүүлэлтийн зарчим нь L матрицын мөр бүр дээрх факторуудын ачилтын квадратуудын

дисперсийг максимумчлах өөрөөр хэлбэл эргүүлэлттэй ачилт болох L^* матрицын багана бүр дэх q^*_{ij} ачилтын квадратуудын дисперсүүдийн нийлбэр буюу

$$\frac{1}{p}\sum_{i=1}^{k}\left[\sum_{j=1}^{p}(\tilde{q}^*_{ji})^4 - \left\{\frac{1}{p}\sum_{j=1}^{p}(\tilde{q}^*_{ji})^2\right\}^2\right]$$

хэмжигдэхүүнийг максимумчлах байдлаар J эргүүлэлтийг олж тогтоодог. Энд $\tilde{q}^*_{ji} = q^*_{ji}/h^*_j$ байна. Иймд тус эргүүлэлтийг хувьсагчийн факторууд дээрх ялгарлыг тодруулдаг гэж болно.

Эргүүлэлттэй үнэлгээ

Хэрэв эргүүлэлт нь J матрицаар тодорхойлогдож байвал факторын ачилтыг

$$\hat{L}^* = \hat{L}J$$

харин факторын оноог

$$\hat{F}^* = J^T\hat{F}$$

байдлаар хувиргах буюу эргүүлнэ. `psych::principal()` функцийн хувьд эргүүлэлтийг `rotate` аргументын тусламжтай заана.

```
psych::principal(r = X, nfactors = k, rotate = "varimax")
```

Тус функцийн хувьд эргүүлэлт, эргүүлэлтийн матриц, эргүүлэлттэй ачилт, эргүүлэлттэй оноо зэргийг түүний буцаах утга доторх `rotation`, `rot.mat`, `loadings`, `scores` элементүүдэд оноож өгсөн байдаг.

Лекц X
Кластерын шинжилгээ I

1 Кластерын шинжилгээ

Кластерын шинжилгээ

Кластерын шинжилгээ нь юмс үзэгдэл дээрх ажиглалт болох олон хэмжээст өгөгдөл эсвэл түүнээс зохиосон зайн матриц дээр тулгуурлан юмс үзэгдлийн ангилал, тусгаар байдлыг тогтоодог. Шинжилгээний зорилго бол өөр бүлгүүдэд байх юмс үзэгдлээс илүү адил төстэй юмс үзэгдлийг нэг бүлэгт бүлэглэх явдал юм. Тэрхүү бүлгийг *кластер* гэдэг. Үүнийг нөгөө талаас нь харвал ялгаатай юмс үзэгдлийг ангилан ялгах байдлаар кластер гэх бүлгүүдэд хуваарилж буй явдал юм.

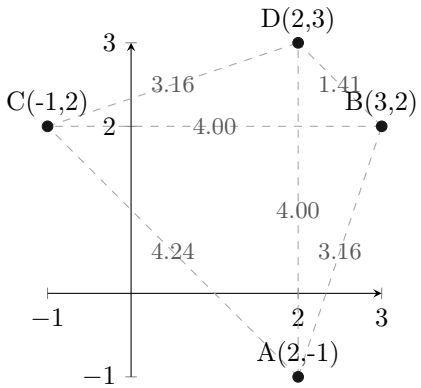
Жишээ 32. A, B, C, D дөрвөн объектыг $X = (X_1, X_2)$ векторын утгуудаар ангилсан кластерын шинжилгээ хий.

```
X <- matrix(data = c(2,-1,3,2,-1,2,2,3), nrow = 4, byrow = TRUE,
    dimnames = list(LETTERS[1:4], c("X1", "X2")))
```

```
d <- dist(X, method = "euclidean")
```

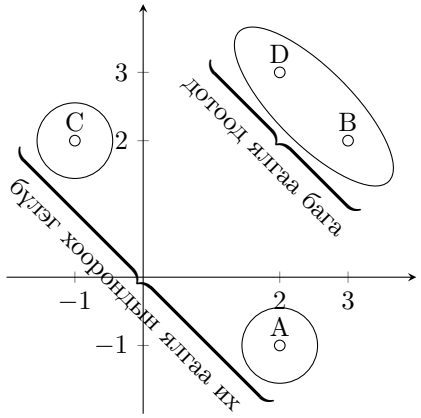
Объект	X_1	X_2
A	2	−1
B	3	2
C	−1	2
D	2	3

Хүснэгт 3: Кластерын шинжилгээнд ашиглах өгөгдөл



Зураг 16: Объектууд буюу цэгүүд, тэдгээрийн хоорондох Евклидийн зай

Кластерын шинжилгээний үндсэн зарчим



Зураг 17: Кластер байгуулах үндсэн зарчим

2 Кластер байгуулах алгоритмуудаас

Кластер байгуулах алгоритмуудаас

- Шатлах (hierarchical clustering) Объект хоорондын зай дээр тулгуурладаг.
  - Цуглуулах арга Шинжилгээний эхэнд объект бүрийг тусдаа кластер гэж үзэх ба улмаар тэдгээр кластеруудаа бүлэглэж нэгтгэдэг.
  - Хуваах арга Шинжилгээний эхэнд бүх объект нэг кластерт байна гэж үзэх ба улмаар тэдгээрийг задалж олон кластерт ангилдаг.

  Кластеруудын ялгаа буюу тэдгээрийн хоорондох зайг объект хоорондын зайнаас тооцож гаргадаг.

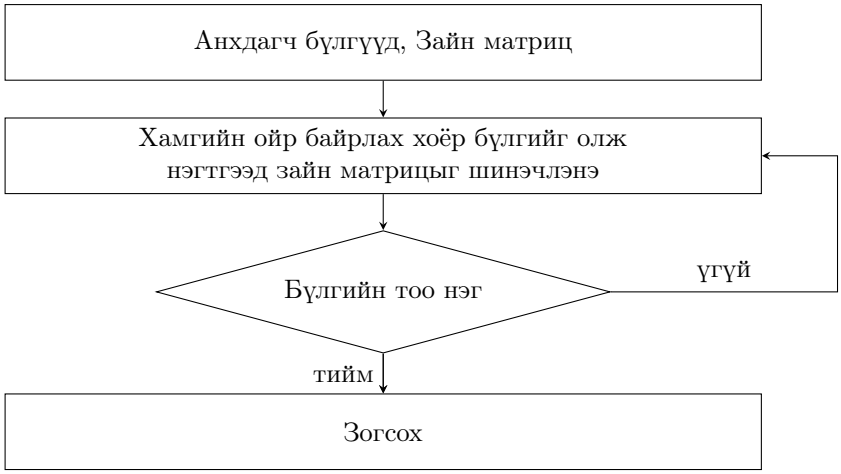
- Хэсэгчлэх (k-means clustering) Заасан бүлгийн тоонд харгалзах түр зуурын бүлэглэлтээс эхэлж улмаар тооцооны үр дүнг оптимал болтол элементүүдийг бүлгүүдийн хооронд сэлгэж шилжүүлэх байдлаар ангиллын оптимал шийд олох буюу кластерууд байгуулдаг. Зай тооцохдоо кластерыг түүний төв буюу дунджаар нь төлөөлүүлдэг.

3 Шатлах алгоритм

Шатлах алгоритмаар хийх кластерын шинжилгээнд шаардагдах зүйлс

1. Зайн матриц Объектуудын хоорондох зай буюу ялгаатай байдлыг илэрхийлсэн матриц
2. Кластер хоорондын зай тооцох аргачлал Алгоритм ажиллах явцад үүсэх шинэ кластер ба бусад кластер хоорондын зай тооцоолох арга замыг заасан дүрэм

Шатлах алгоритмын цуглуулах аргын ажиллах зарчим



Зайн матриц олох буюу зайн хэмжээс сонгох тухай

Хоёр объектын ялгаатай байдлыг зөв хэмжих буюу тохирох зайн хэмжээс сонгож хэрэглэх нь тус шинжилгээний эцсийн үр дүнд их нөлөөтэй. Зайн хэмжээс нь санамсаргүй хувьсагчдын хэмжээсийн төрлөөс, цаашилбал тоон хувьсагчийн хувьд түүний утга агуулгаас, чанарын хувьсагчийн хувьд ангийнх нь тоо болон давтамж зэргээс хамаарна. Тоон хувьсагчийн хувьд зайн хэмжээсүүд ялгаатай байдлыг шууд илэрхийлдэг байхад чанарын хувьсагчийн хувьд эсрэгээрээ адил төстэй байдлыг хэмжсэний дараа түүнийгээ урвуулж ялгаатай байдлын илэрхийлэл болгодог.

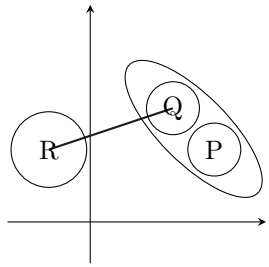
Кластер хоорондын зай тооцох аргачлалуудаас

Томьёо 3 ($P+Q$ шинэ бүлэг ба R бүлэг хоорондын зай)**.**

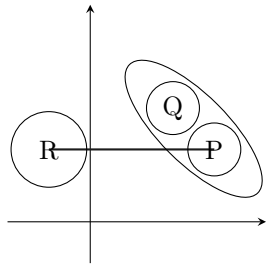
$$\mathrm{d}(R, P+Q) = \sigma_1\,\mathrm{d}(R,P) + \sigma_2\,\mathrm{d}(R,Q) + \sigma_3\,\mathrm{d}(P,Q) + \sigma_4|\,\mathrm{d}(R,P) - d(R,Q)|$$

Зай	σ_1	σ_2	σ_3	σ_4
Single linkage	$1/2$	$1/2$	0	$-1/2$
Complete linkage	$1/2$	$1/2$	0	$1/2$
Average linkage	$1/2$	$1/2$	0	0
Average linkage[12]	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroid	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Median	$1/2$	$1/2$	$-1/4$	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

Single linkage алгоритм



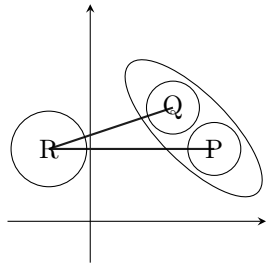
$$\begin{aligned}\mathrm{d}(R, P+Q) &= \frac{1}{2}\,\mathrm{d}(R,P) + \frac{1}{2}\,\mathrm{d}(R,Q) - \frac{1}{2}|\,\mathrm{d}(R,P) - d(R,Q)| \\ &= \min\{\mathrm{d}(R,P), \mathrm{d}(R,Q)\}\end{aligned}$$



Complete linkage алгоритм

$$\begin{aligned} \mathrm{d}(R, P+Q) &= \frac{1}{2}\,\mathrm{d}(R,P) + \frac{1}{2}\,\mathrm{d}(R,Q) + \frac{1}{2}|\,\mathrm{d}(R,P) - d(R,Q)| \\ &= \max\{\mathrm{d}(R,P), \mathrm{d}(R,Q)\} \end{aligned}$$

Average linkage алгоритм



- жинлээгүй

$$\mathrm{d}(R, P+Q) = \frac{\mathrm{d}(R,P) + \mathrm{d}(R,Q)}{2}$$

- жинлэсэн

$$\mathrm{d}(R, P+Q) = \frac{n_P\,\mathrm{d}(R,P) + n_Q\,\mathrm{d}(R,Q)}{n_P + n_Q}$$

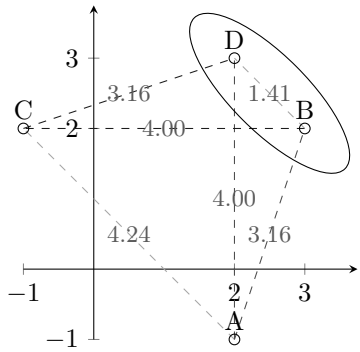
```
hc <- hclust(d, method = "average")
plot(hc)
hc$height
```

Centroid болон Median алгоритмууд

$$\mathrm{d}(R, P+Q) = \frac{n_P}{n_P+n_Q}\,\mathrm{d}(R,P) + \frac{n_Q}{n_P+n_Q}\,\mathrm{d}(R,Q) - \frac{n_P n_Q}{(n_P+n_Q)^2}\,\mathrm{d}(P,Q)$$

$n_p = n_Q = n_R = 1$ үед Median алгоритм болно. Энэ тохиолдолд зайг Евклидийн нормоор хэмжсэн үед кластер хоорондын зай нь гурвалжны медиан гэсэн геометр утга агуулгатай болдог.

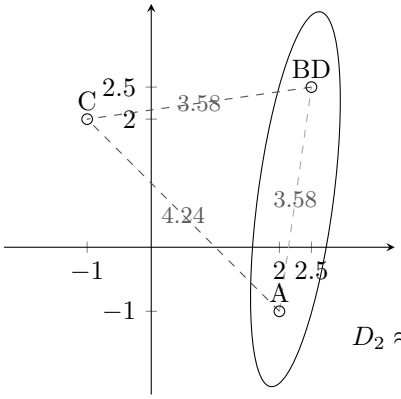
Жинлэсэн Average linkage алгоритм, I итерац



$$D \approx \begin{pmatrix} 0 & 3.16 & 4.24 & 4.00 \\ & 0 & 4.00 & \boxed{1.41} \\ & & 0 & 3.16 \\ & & & 0 \end{pmatrix}$$

$$D_1 \approx \begin{pmatrix} 0 & \frac{3.16+4.00}{2} = 3.58 & 4.24 \\ & 0 & 3.58 \\ & & 0 \end{pmatrix}$$

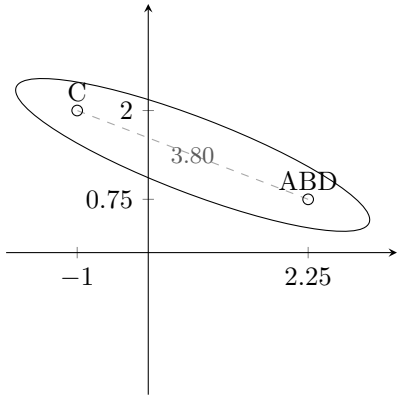
Жинлэсэн Average linkage алгоритм, II итерац



$$D_1 \approx \begin{pmatrix} 0 & \boxed{3.58} & 4.24 \\ & 0 & 3.58 \\ & & 0 \end{pmatrix}$$

$$D_2 \approx \begin{pmatrix} 0 & \frac{1}{3} \cdot 4.24 + \frac{2}{3} \cdot 3.58 = 3.80 \\ & 0 \end{pmatrix}$$

Жинлэсэн Average linkage алгоритм, III итерац

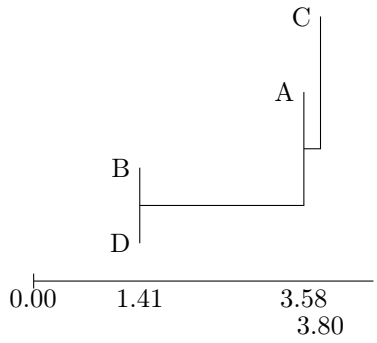


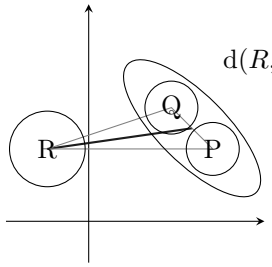
$$D_2 \approx \begin{pmatrix} 0 & \boxed{3.80} \\ & 0 \end{pmatrix}$$

Ward алгоритм

Энэ нь нэгтгэсний дараа дотоод ялгаа нь хэт ихсэхгүй байх бүлгүүдийг олж нэгтгэх зарчимтай алгоритм юм. Бүлгийн дотоод ялгаа дараах томьёогоор

Жинлэсэн Average linkage алгоритмаар хийсэн кластерын шинжилгээний үр дүнг харуулсан дендрограмм





$$\mathrm{d}(R, P+Q) = \sqrt{\frac{\mathrm{d}^2(R,P)}{2} + \frac{\mathrm{d}^2(R,Q)}{2} - \frac{\mathrm{d}^2(P,Q)}{4}}$$

$\mathrm{d}(R, P+Q)$ бол $\triangle PRQ$ гурвалжны медиан буюу R ба $P+Q$ бүлгийн төв хоорондын зай

илэрхийлэгдэнэ.

$$D_R = \frac{1}{n_R} \sum_{i=1}^{n_R} \mathrm{d}^2(x_i, \overline{x}_R)$$

Энд $\overline{x}$ нь бүлгийн төв юм. Евклидийн зай ашигласан үед бүлгийн төв нь дундаж, бүлгийн дотоод ялгаа нь дисперс зэрэг статистикуудтай давхацна. Хоёр бүлэг нэгдэхэд гарах дотоод ялгааны өөрчлөлт дараах томьёогоор илэрхийлэгдэнэ.

$$\Delta(P,Q) = \underbrace{D_P + D_Q}_{\text{өмнөх нийт дотоод ялгаа}} - \underbrace{D_{P+Q}}_{\text{дараах дотоод ялгаа}} = \frac{n_P n_Q}{n_P + n_Q}\,\mathrm{d}^2(P,Q)$$

Шатлах алгоритмаар хийх кластерын шинжилгээний `hclust()` функц

```
hclust(d, method = "complete")
```

`d` зайн матриц

`method` кластер хоорондын зай бодох аргачлал: `"single"`, `"complete"`, `"average"`, `"median"`, `"centroid"`, `"ward.D"`

Функцийн буцаах утгын тайлбар

merge шатлах алгоритмын алхам бүрт ямар хоёр кластер нэгдсэнийг илтгэх хоёр баганатай матриц

height шатлах алгоритм ажиллах явцад шинэ кластер байгуулагдах үеийн кластер хоорондын зайн утга

order объект буюу цэгүүдийг дендрограмм байгуулахад зохимжтой байдлаар сонгох дэс дараалал

labels объект буюу цэгүүдийн нэр

method кластер хоорондын зай бодох аргачлал

dist.method `d` матрицын `method` аттрибутын утга буюу зайн хэмжээс

Лекц XI
Кластерын шинжилгээ II

1 Хэсэгчлэх алгоритм

Кластерын шинжилгээний хэсэгчлэх алгоритм

Сэргээн санах нь 7. Хэсэгчлэх алгоритм (k-means clustering) нь заасан бүлгийн тоонд харгалзах түр зуурын бүлэглэлтээс эхэлж улмаар тооцооны үр дүнг оптимал болтол элементүүдийг бүлгүүдийн хооронд сэлгэж шилжүүлэх байдлаар ангиллын оптимал шийд олох буюу кластерууд байгуулдаг. Зай тооцохдоо кластерыг түүний төв буюу дунджаар нь төлөөлүүлдэг.

Объект	X_1	X_2
A	5	3
B	−1	1
C	1	−2
D	−3	−2

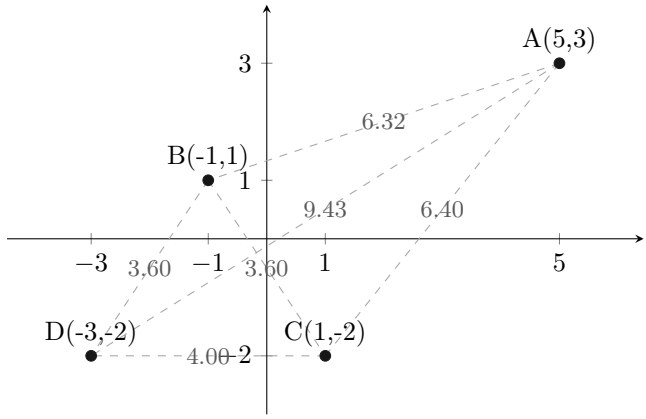
Хүснэгт 4: Кластерын шинжилгээнд ашиглах өгөгдөл

Жишээ 33. A, B, C, D дөрвөн объектыг $X = (X_1, X_2)$ векторын утгуудаар ангилсан кластерын шинжилгээ хий.

```
X <- matrix(data = c(5,3,-1,1,1,-2,-3,-2), nrow = 4, byrow = TRUE)
print(X)
```

Хэсэгчлэх алгоритмаар хийх кластерын шинжилгээний `kmeans()` функц

```
kmeans(x, centers)
```



Зураг 18: Объектууд буюу цэгүүд, тэдгээрийн хоорондох Евклидийн зай

`x` өгөгдөл

`centers` кластерын тоо

```
kmeans(x = X, centers = 2)
```

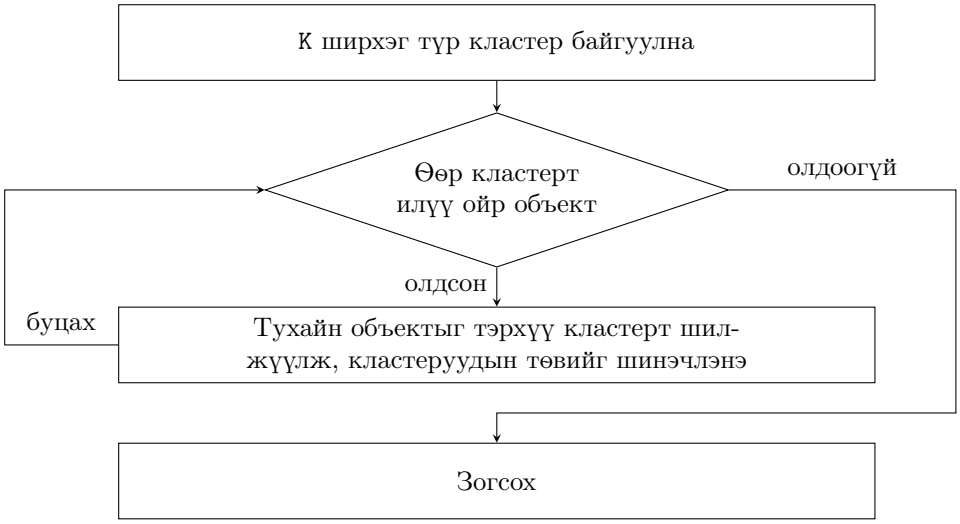
```
K-means clustering with 2 clusters of sizes 3, 1

Cluster means:
  [,1] [,2]
1   -1   -1
2    5    3

Clustering vector:
[1] 2 1 1 1

Within cluster sum of squares by cluster:
[1] 14  0
 (between_SS / total_SS =  73.6 %)

Available components:

[1] "cluster"     "centers"   "totss"    "withinss"   "tot.withinss"
[6] "betweenss"   "size"      "iter"     "ifault"
```

Кластерын шинжилгээний хэсэгчлэх алгоритм

2 Кластерын төв

Кластерын төв

Кластерын төв нь тархалтын төвийн үзүүлэлтээр шууд тодорхойлогдоно. Иймд түүнийг түүврийн дунджаар үнэлнэ.

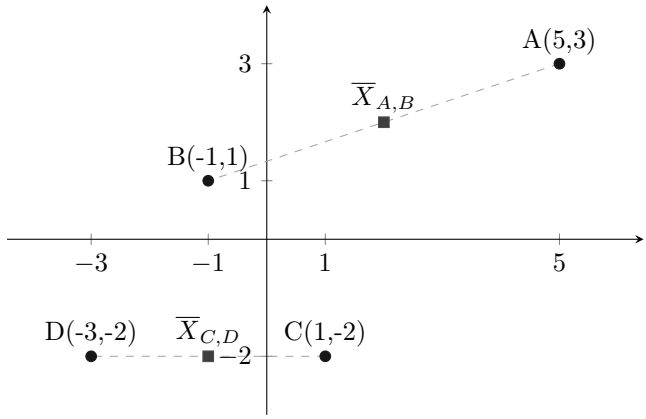


Томьёо 4 (Кластерын төв шинэчлэх)**.** X объектыг нэг кластераас нөгөөд шилжүүлэх үед

$$\overline{X}_j := \begin{cases} \dfrac{n\overline{X}_j + X_j}{n+1} & \text{объект кластерт нэгдэн орсон} \\ \dfrac{n\overline{X}_j - X_j}{n-1} & \text{объект кластераас гарсан} \end{cases}$$

$\overline{X}_j$ нь кластерын төвийн j дүгээр координат, X_j нь X объектын j дүгээр координат, n нь "хуучин" кластерын хэмжээ

Түр кластерын төв: $\{A, B\}$ **ба** $\{C, D\}$



$$\overline{X}_{A,B} = \left(\frac{5+(-1)}{2}, \frac{3+1}{2}\right) = (2,2) \qquad \overline{X}_{C,D} = (-1,-2)$$

Кластерын төв шинэчлэх: $\{A\}$ **ба** $\{B, C, D\}$

В цэгийн координат $(-1, 1)$ болохыг анхаарч кластерын төв шинэчлэх томьёо ашиглавал

- хуучин $\{A, B\}$ кластерын хэмжээ $n = 2$ бөгөөд түүний төв $\overline{X}_{A,B} = (2, 2)$ байсан тул шинэ $\{A\}$ кластерын төвийн координатууд

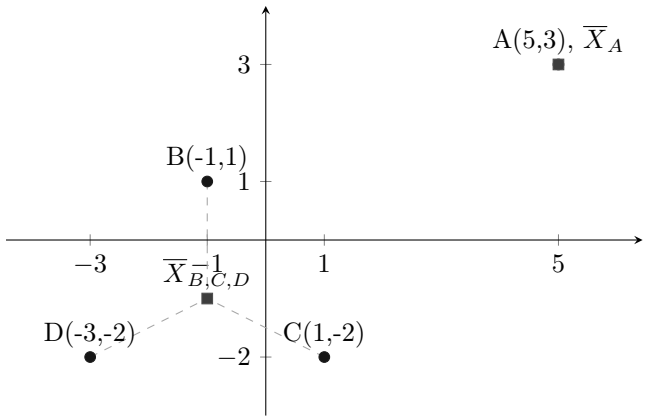
$$\overline{X}_1 = \frac{2 \cdot (2) - (-1)}{2 - 1} = 5 \qquad \overline{X}_2 = \frac{2 \cdot (2) - 1}{2 - 1} = 3$$

- хуучин $\{C, D\}$ кластерын хэмжээ $n = 2$ бөгөөд түүний төв $\overline{X}_{C,D} = (-1, -2)$ байсан тул шинэ $\{B, C, D\}$ кластерын төвийн координатууд

$$\overline{X}_1 = \frac{2 \cdot (-1) + (-1)}{2 + 1} = -1 \qquad \overline{X}_2 = \frac{2 \cdot (-2) + 1}{2 + 1} = -1$$

гэж олдоно.

Шинэчлэгдсэн кластерын төв: $\{A\}$ **ба** $\{B, C, D\}$



$$\overline{X}_{B,C,D} = \left(\frac{-1 + 1 + (-3)}{3}, \frac{1 + (-2) + (-2)}{3} \right) = (-1, -1)$$

3 Өөр кластерт илүү ойр объект олох нь

Өөр кластерт илүү ойр объект хайх нь

Дараах тооцоог объект нэг бүрчлэн хийж, өөр кластерт шилжүүлбэл зохих объектыг олно. Энд d^2 нь зайн квадрат юм.

Кластерууд	Тухайн объектыг шилжүүлээгүй үед	Тухайн объектыг шилжүүлсэн үед
Кластер №1	d^2(объект, кластерын төв)	d^2(объект, төв)
⋮	⋮	⋮
Кластер №K	d^2(объект, төв)	d^2(объект, төв)

Өөр кластерт илүү ойр объект хайх: I итерац
$A = (5, 3)$ *объект*

Шилжүүлээгүй үед

$$d^2(A, \{A, B\}) = d^2((5, 3), (2, 2)) = (5 - 2)^2 + (3 - 2)^2 = \boxed{10}$$
$$d^2(A, \{C, D\}) = d^2((5, 3), (-1, -2)) = (5 - (-1))^2 + (3 - (-2))^2 = 61$$

Шилжүүлсэн үед

$$d^2(A, \{B\}) = d^2((5, 3), (-1, 1)) = (5 - (-1))^2 + (3 - 1)^2 = 40$$
$$d^2(A, \{A, C, D\}) = d^2((5, 3), (1, -1/3)) = (5 - 1)^2 + (3 - (-1/3))^2 \approx 27.09$$

$\min d^2 = 10 = d^2(A, \{A, B\})$ тул A объектыг шилжүүлэх шаардлагагүй.

Өөр кластерт илүү ойр объект хайх: I итерац
$B = (-1, 1)$ *объект*

Шилжүүлээгүй үед

$$d^2(B, \{A, B\}) = d^2((-1, 1), (2, 2)) = (-1 - 2)^2 + (1 - 2)^2 = 10$$
$$d^2(B, \{C, D\}) = d^2((-1, 1), (-1, -2)) = (-1 - (-1))^2 + (1 - (-2))^2 = 9$$

Шилжүүлсэн үед

$$d^2(B, \{A\}) = d^2((-1, 1), (5, 3)) = (-1 - 5)^2 + (1 - 3)^2 = 40$$
$$d^2(B, \{B, C, D\}) = d^2((-1, 1), (-1, -1)) = (-1 - (-1))^2 + (1 - (-1))^2 = \boxed{4}$$

$\min d^2 = 4 = d^2(B, \{B, C, D\})$ тул B объектыг шилжүүлэх боломжтой.

Өөр кластерт илүү ойр объект хайх: I итерац
C болон D объект

- C объект

$$d^2(C, \{A, B\}) = 17 \quad d^2(C, \{A, B, C\}) \approx 7.55$$
$$d^2(C, \{C, D\}) = \boxed{4} \quad d^2(C, \{D\}) = 16$$

  C объектыг шилжүүлэх шаардлагагүй.

- D объект

$$d^2(D, \{A, B\}) = 41 \quad d^2(D, \{A, B, D\}) \approx 18.22$$
$$d^2(D, \{C, D\}) = \boxed{4} \quad d^2(D, \{C\}) = 16$$

  D объектыг шилжүүлэх шаардлагагүй.

I итерацийн үр дүн $d^2(B, \{B, C, D\}) = 4$ нь шилжүүлэх боломжтой бүх тохиолдол дундаас хамгийн бага нь тул B объектыг $\{C, D\}$ кластерт шилжүүлнэ.

Өөр кластерт илүү ойр объект хайх: II итерац
$C = (1, -2)$ *объект*

Шилжүүлээгүй үед

$$d^2(C, \{A\}) = d^2((1, -2), (5, 3)) = (1 - 5)^2 + (-2 - 3)^2 = 41$$
$$d^2(C, \{B, C, D\}) = d^2((1, -2), (-1, -1)) = (1 - (-1))^2 + (-2 - (-1))^2 = \boxed{5}$$

Шилжүүлсэн үед

$$d^2(C, \{A, C\}) = d^2((1, -2), (3, 0.5)) = (1 - 3)^2 + (-2 - 0.5)^2 = 10.25$$
$$d^2(C, \{B, D\}) = d^2((1, -2), (-2, -0.5)) = (1 - (-2))^2 + (-2 - (-0.5))^2 = 11.25$$

$\min d^2 = 5 = d^2(C, \{B, C, D\})$ тул C объектыг шилжүүлэх шаардлагагүй.

Өөр кластерт илүү ойр объект хайх: II итерац
$D = (-3, -2)$ *объект*

Шилжүүлээгүй үед

$$d^2(D, \{A\}) = d^2((-3, -2), (5, 3)) = (-3 - 5)^2 + (-2 - 3)^2 = 99$$
$$d^2(D, \{B, C, D\}) = d^2((-3, -2), (-1, -1)) = (-3 - (-1))^2 + (-2 - (-1))^2 = \boxed{5}$$

Шилжүүлсэн үед

$$d^2(D, \{A, D\}) = d^2((-3, -2), (1, 0.5)) = (-3 - 1)^2 + (-2 - 0.5)^2 = 22.25$$
$$d^2(D, \{B, C\}) = d^2((-3, -2), (0, -0.5)) = (-3 - 0)^2 + (-2 - (-0.5))^2 = 11.25$$

$\min d^2 = 5 = d^2(D, \{B, C, D\})$ тул D объектыг шилжүүлэх шаардлагагүй.

Өөр кластерт илүү ойр объект хайх: II итерац
A болон B объект

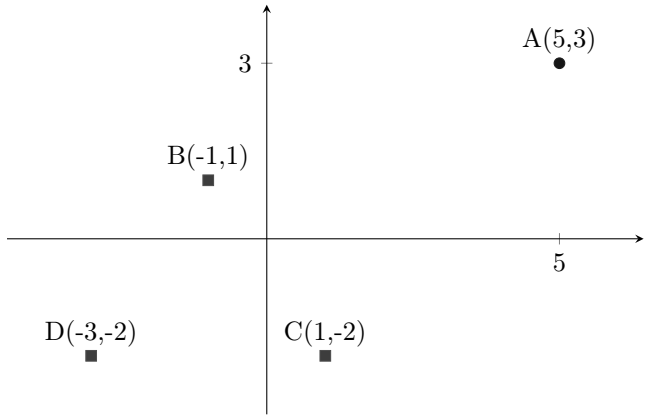
A объектыг шилжүүлбэл $\{A\}$ кластер хоосон болох тул үүнийг алгасна. B объект $\{A\}$ кластерт бус харин $\{C, D\}$ кластерт илүү ойр болох нь өмнөх итерацаар тогтоогдсон тул үүнд харгалзах тооцоог алгасна. **II итерацийн үр дүн** Өөр кластерт шилжүүлж болох объект олдсонгүй. Иймд алгоритмын дагуу бодолтыг зогсооно.

Эцсийн үр дүн: Кластер хуваалт

Зайн квадратуудын нийлбэр

Кластерын дотоод зайн квадратуудын нийлбэр SSI

$$\begin{aligned} &SS_{\{B,C,D\}} \\ &\quad = d^2(B, \{B, C, D\}) + d^2(C, \{B, C, D\}) + d^2(D, \{B, C, D\}) \\ &\quad = 4 + 5 + 5 = 14 \\ &SS_{\{A\}} = d^2(A, \{A\}) = 0 \\ &SSI = SS_{\{B,C,D\}} + SS_{\{A\}} = 14 + 0 = 14 \end{aligned}$$



Нийт зайн квадратуудын нийлбэр SST

$$SST = \sum_{i=1}^{n} d^2(X_i, \overline{X}) = 29.25 + 3.25 + 4.25 + 16.25 = 53$$

Энд X_i нь түүврийн элемент, $\overline{X}$ нь түүврийн дундаж юм.

Кластер хоорондын зайн квадратуудын нийлбэр SSB

$$SSB = \sum_{i=1}^{K} n_i d^2(\overline{X}_i, \overline{X}) = 1 \cdot 29.25 + 3 \cdot 3.25 = 39$$

Энд $\overline{X}_i$ нь кластерын төв, n_i нь кластерын хэмжээ, $\overline{X}$ нь кластеруудын төв юм.

Чанар 13.

$$SST = SSI + SSB$$

$$\frac{SSB}{SST} \cdot 100\% = \frac{39}{53} \cdot 100\% \approx 73.58\%$$

4 Хувьсагчдын масштабын ялгаатай байдлын нөлөөг зайлуулах

Хувьсагчдын масштабын ялгаатай байдлын нөлөөг зайлуулах

Хувьсагчдын хэмжээс тухайлбал хэмжүүрийн нэгж ялгаатай үед зарим хувьсагч зайн хэмжээст хэт давамгайлах байдал үүсэх талтай. Улмаар энэ нь кластерын шинжилгээний үр дүнг гажуудуулна. Ийм тохиолдолд хувьсагчдын масштабыг тэгшитгэхийн тулд A метриктэй Евклидийн

$$d_{ij}^2 = ||x_i - x_j||_A = (x_i - x_j)^T A (x_i - x_j)$$

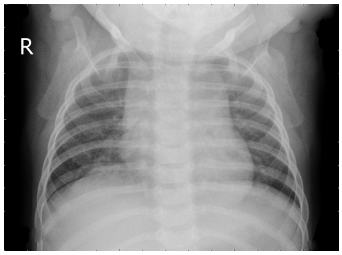
норм ашиглана. Хэрэв масштаб хэрхэн тэгшитгэх нь тодорхойгүй бол $A = \text{diag}(s_{X_1X_1}^{-1}, \ldots, s_{X_pX_p}^{-1})$ буюу хувьсагчдын дундаж квадрат хазайлт ашиглаж болно. Угтаа энэ нь хувьсагчдыг стандарт хувиргалтаар хувиргасантай адил болох тул тухайлбал R програм дээр өгөгдлийн матрицаа `X <- scale(X)` байдлаар хувиргана.

Лекц XII
Дискриминантын шинжилгээ

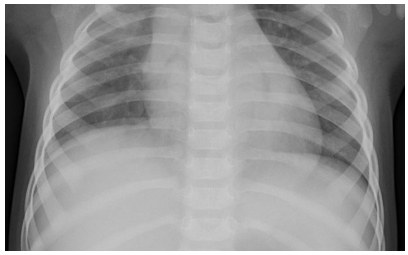
1 Дискриминантын шинжилгээ

Дискриминантын шинжилгээгээр шийдэж болох нэг асуудал

Жишээ 34. www.kaggle.com/paultimothymooney/chest-xray-pneumonia веб хуудас дээрх уушгины хатгааны рентген зургийн мэдээлэлд үндэслэн эрүүл ба хатгаатай хүмүүсийг ялгах машин сургалтын загвар боловсруул.



(a) эрүүл



(b) бактерын гаралтай хатгаа

Зураг 19: Цээжний рентген зураг

Дискриминантын шинжилгээг хэдийд хийж болох тухай

$$\underbrace{\text{чанарын хувьсагч}}_{\text{хамааран хувьсагч}} \sim \underbrace{\text{тоон хувьсагч}}_{\text{үл хамааран хувьсагч}}$$

```
MASS::lda(formula = type ~ ., data = chest_xray)
```

`type` эрүүл ба хатгаатай эсэхийг заасан фактор хэлбэртэй чанарын хувьсагч

`. type` хувьсагчаас бусад хувьсагч

`chest_xray` өгөгдөл агуулж буй датафрейм

Өгөгдөл бэлдэх

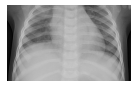
IM-0115-0001.jpeg

IM-0117-0001.jpeg

person1_bacteria_1.jpeg

person100_virus_184.jpeg

$$\text{зураг} \xrightarrow{\text{RGB}} \underbrace{\text{grayscale}}_{\text{матриц}} \rightarrow \text{SVD} \rightarrow \underbrace{\text{сингуляр буюу хувийн утгууд}}_{\text{вектор}}$$

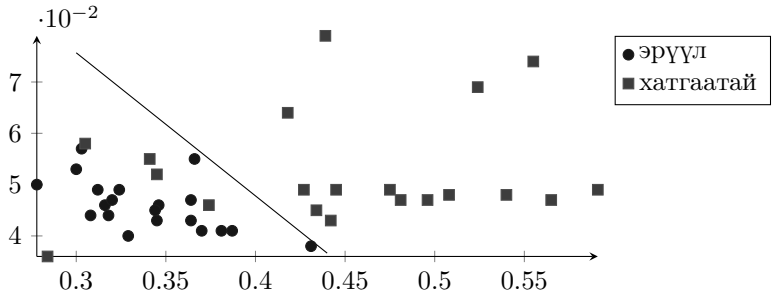


$\longrightarrow \quad (0.565, 0.047, \text{bacteria})$

№	SV1	SV2	type
1	0.3664534	0.05499858	normal
21	0.5650811	0.04686256	bacteria
41	0.4338590	0.04457886	virus

Хүснэгт 5: Бэлдсэн өгөгдлийн зарим мөр

Дискриминантын шинжилгээний үр дүн



Зураг 20: Дискриминантын шинжилгээний үр дүн

Дискриминантын шинжилгээний тухай

Дискриминантын шинжилгээ бол эд юмс болон үзэгдлийг түүний шинж чанарыг илтгэх хувьсагчдын тусламжтай ялгаж таних болон ангилж тусгаарлах арга загварыг судалдаг олон хэмжээст өгөгдлийн статистик шинжилгээний нэг чиглэл юм. Орчин үед тус шинжилгээг хэв танилт, машин сургалт зэрэгт өргөн хэрэглэж байна. Бид энэ удаагийн хичээлээр шугаман дискриминантын шинжилгээний талаар түлхүү үзнэ. Шугаман дискриминантын шинжилгээ нь машин сургалтын Support Vector Machine аргын статистик аналог юм.

Шугаман дискриминантын шинжилгээ нь статистикийн бусад арга, загвартай ч нягт уялдаатай.

- **Кластерын шинжилгээ** Юмсийг ангилах зорилгоороо ижил боловч анги, бүлгийг урьдчилж заах эсвэл заахгүйгээрээ ялгаатай.

- **Логистик регресс** Чанарын хувьсагчийг хамааран хувьсагч болгон авдагаараа бас чанарын хувьсагчийн утгыг прогнозлох буюу шинж төлвийг нь тогтооход ашигладагаараа төстэй юм.
- **Гол хэсгийн шинжилгээ ба Факторын шинжилгээ** Эдгээр нь өгөгдлийг хамгийн сайн тайлбарлаж чадах шугаман эвлүүлэг буюу шулуун хайдаг. Тухайлбал гол хэсгийн шинжилгээ хамгийн их дисперстэй чиглэлд харгалзах шулуун олдог бол шугаман дискриминантын шинжилгээ нь бүлгүүдийг хамгийн сайн зааглаж тусгаарлах шулуун олдог.

2 Ангиллын зарчим

Дискриминантын шинжилгээний ангиллын зарчим

X санамсаргүй векторын тодорхой нэг утгыг төлөөлөх x цэгийг Π_j бүлэг буюу эх олонлогт харьяалуулах эсэхийг шийдэх дүрмийг *ангиллын зарчим* гэнэ. Дискриминантын шинжилгээнд дараах нэр бүхий ангиллын зарчмууд байдаг.

- Хамгийн их үнэний хувь бүхий дискриминантын зарчим
- Байесийн дискриминантын зарчим
- Фишерийн шугаман дискриминантын зарчим

Ангиллын зарчмаар Π_j эх олонлогт харьяалагдах цэгүүдийн олонлог буюу ангиллын мужийг R_j өөрөөр хэлбэл

$$R_j = \{x : x \in \Pi_j\}$$

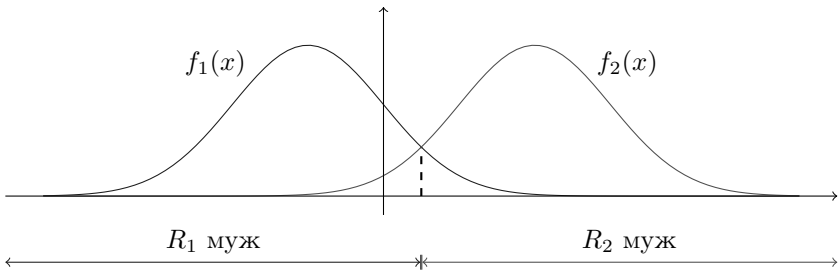
гэе.

Хамгийн их үнэний хувь бүхий дискриминантын зарчим

x цэгийн хувьд

$$L_j(x) = \max\{L_1(x), L_2(x)\} \text{ буюу } f_j(x) = \max\{f_1(x), f_2(x)\}$$

бол түүнийг Π_j эх олонлогт хуваарилна.



Зураг 21: Хамгийн их үнэний хувь бүхий дискриминантын зарчим

$$R_1 = \{x : f_1(x) \geq f_2(x)\} \qquad R_2 = \{x : f_2(x) > f_1(x)\}$$

Жишээ 35. Санамсаргүй хувьсагчийн эх олонлогуудын тархалт $N_1(\mu_j, \sigma_j^2)$ бол хамгийн их үнэний хувь бүхий дискриминантын зарчимд харгалзах ангиллын мужуудыг ол.

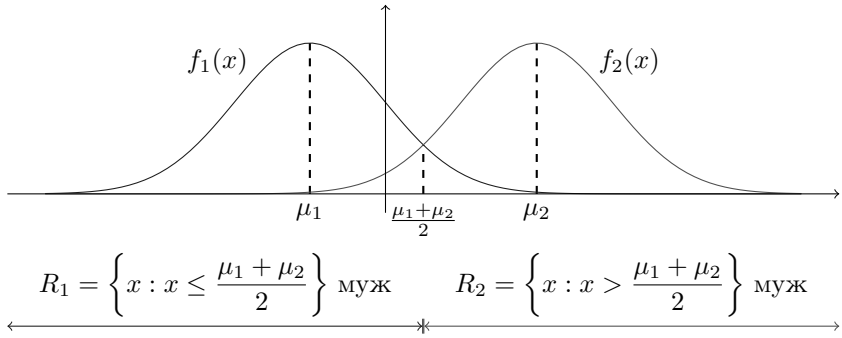
$R_1 = \{x : f_1(x) \geq f_2(x)\}$ мужийг тодорхойлох нөхцөл

$$f_1(x) \geq f_2(x)$$
$$\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \geq \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}$$
$$x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) \leq 2\ln\frac{\sigma_2}{\sigma_1}$$

болно. Ийнхүү $\sigma_1 = \sigma_2$ тухайн тохиолдолд $\mu_1 < \mu_2$ үед

$$R_1 = \left\{x : x \leq \frac{\mu_1 + \mu_2}{2}\right\} \quad R_2 = \left\{x : x > \frac{\mu_1 + \mu_2}{2}\right\}$$

мужууд олдоно.



Зураг 22: Ижил дисперстэй хэвийн тархалттай эх олонлогуудын хувьд хамгийн их үнэний хувь бүхий дискриминантын зарчмаар тодорхойлогдох ангиллын мужууд

Алдаатай ангиллын хохирол тооцсон ангиллын муж

x цэгийг алдаатай ангилсанаас учрах хохирлын хэмжээг C гэе.

$$EC \to \min \;\Rightarrow\; \exists R_j$$

	Π_1	Π_2
R_1	✓	×
R_2	×	✓

(a) зөв ба буруу ангилал

	Π_1	Π_2
R_1	0	$C(1\|2)$
R_2	$C(2\|1)$	0

(b) хохирлын хэмжээ

Хүснэгт 6: Зөв ба буруу ангилал, түүнээс үүдэн учрах хохирлын хэмжээ

$$\begin{aligned} EC &= C(1|2)\cdot P(R_1\Pi_2) + C(2|1)\cdot P(R_2\Pi_1) + 0\cdot P(R_1\Pi_1 + R_2\Pi_2) \\ &= C(1|2)\cdot P(R_1|\Pi_2)\cdot P(\Pi_2) + C(2|1)\cdot P(R_2|\Pi_1)\cdot P(\Pi_1) \\ &= C(1|2)\cdot P(R_1|\Pi_2)\cdot \pi_2 + C(2|1)\cdot P(R_2|\Pi_1)\cdot \pi_1 \end{aligned}$$

Энд $\pi_j = P(\Pi_j) = P(x \in \Pi_j)$ гэж тэмдэглэв. Үүнийг приор магадлал гэдэг.

Теорем 9. *EC* буюу буруу ангиллаас үүдэх хохирлын хэмжээний дундаж

$$R_1 = \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2)}{C(2|1)} \frac{\pi_2}{\pi_1} \right\}$$
$$R_2 = \left\{ x : \frac{f_1(x)}{f_2(x)} < \frac{C(1|2)}{C(2|1)} \frac{\pi_2}{\pi_1} \right\}$$

үед хамгийн бага утгадаа хүрнэ.

$\pi_1 = \pi_2$ ба $C(1|2) = C(2|1)$ үед уг зарчим хамгийн их үнэний хувь бүхий зарчимтай давхацна. MASS багцын `lda()` функц π_j приор магадлалуудыг түүвэр дэх бүлгийн хэмжээнд пропорционалаар авдаг бөгөөд хүсвэл өөрөөр заах боломжтой.

$\Pi_j = N_p(\mu_j, \Sigma)$ бас $\pi_1 = \pi_2$ ба $C(1|2) = C(2|1)$ байг.

1. Эх олонлогийн тоо хоёроос их үед: x цэгийг x ба μ_j хоорондын Махаланобисын
$$\delta^2(x, \mu_j) = (x - \mu_j)^T \Sigma^{-1} (x - \mu_j), \quad j = 1 \ldots, J$$
зайг хамгийн бага байлгах Π_j эх олонлогт хуваарилна.

2. Эх олонлогийн тоо хоёртой тэнцүү үед:
$$R_1 = \left\{ x : \alpha^T (x - \mu) \geq 0 \right\}$$
Энд $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ ба $\mu = \dfrac{\mu_1 + \mu_2}{2}$ байна.

Одоо хичээлийн эхэнд авсан жишээгээ үргэлжлүүлье. Эрүүл болон хатгаатай тус бүр 20 хүний цээжний рентген зураг авсан тул приор магадлалуудыг тэнцүү гэж тооцож болно. Бас хувьсагчдыг хоёр хэмжээст хэвийн тархалттай гэе. Тэгвэл өмнөх слайдын 2 дугаар тохиолдол уруу орно. Тэнд эх олонлогуудын ковариацийг адил тэнцүү гэж үзсэн. Иймд ковариацийн матрицыг нийт түүврийн ковариацийн матрицаар үнэлсэн.

$$\hat{\Sigma} = S \approx \begin{pmatrix} 0.00739 & 0.00022 \\ 0.00022 & 0.00008 \end{pmatrix}$$

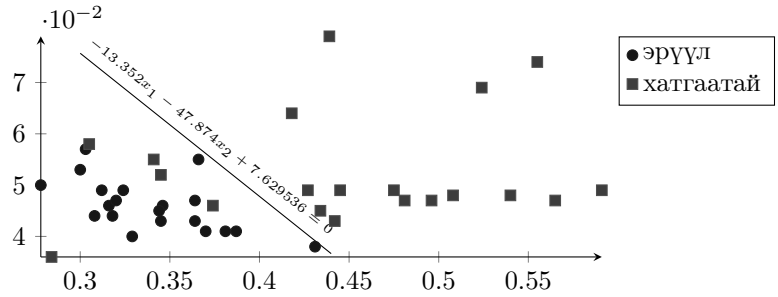
Харин эх олонлогуудын дунджийн үнэлэлт

$$\hat{\mu}_1 \approx (0.340, 0.046) \text{ ба } \hat{\mu}_2 \approx (0.449, 0.053)$$

гэж олдсон.

Ангиллын $R_1 = \left\{ x : \alpha^T (x - \mu) \geq 0 \right\}$ муж дараах хэлбэртэй болно.

$$\alpha^T (x - \mu) \approx (-13.352, -47.874) \begin{pmatrix} x_1 - 0.395 \\ x_2 - 0.049 \end{pmatrix} \approx -13.352 x_1 - 47.874 x_2 + 7.629536 \geq 0$$



Зураг 23: Дискриминантын шинжилгээний үр дүн

MASS багцын `lda()` функцийн зарим үр дүнгийн тайлбар

prior π_j приор магадлалууд

counts бүлэг тус бүрт харгалзах түүврийн хэмжээ

means бүлэг тус бүрийн төв

scaling дискриминантын шулуун дахь хувьсагчдын өмнөх коэффициентүүд

svd бүлгүүдийн хоорондох болон бүлгүүдийн дотоод стандарт хазайлтуудын харьцаа

Постериор магадлал ба хамгийн их постериорын ангилал

$$p_j = P(x \in R_j)$$

буюу статистик загварын тусламжтай зохиосон R_j ангиллын зарчмаар x цэгийн j дүгээр бүлэг буюу эх олонлогт харьяалагдах магадлалыг түүний *постериор магадлал* гэнэ. Мөн x цэгийг хамгийн их постериор магадлалтай бүлэг буюу эх олонлогт хуваарилахыг *хамгийн их постериорын ангилал* гэдэг.

MASS багцын `lda()` функц ашиглаж хийсэн шугаман дискриминантын шинжилгээний хувьд постериор магадлал олохдоо тус функцийн `CV` аргументаар `TRUE` утга дамжуулна. Жишээний хувьд дараах хэлбэртэй код бичнэ.

```
MASS::lda(formula = group ~ ., data = X, CV = TRUE)
```

Постериор магадлал болон хамгийн их постериорын ангилал нь тус функцийн буцаах утгын `posterior` болон `class` элементэд агуулагдана.

Лекц XIII
Олон хэмжээст координатын шинжилгээ

1 Олон хэмжээст координатын шинжилгээний тухай ерөнхий ойлголт

Олон хэмжээст координатын шинжилгээ[13]

Олон хэмжээст координатын шинжилгээ нь их хэмжээст огторгуй дахь цэгүүдийн адил төстэй байдлыг бага хэмжээст огторгуйд тухайлбал хавтгайд буулгаж дүрслэх аргыг судалдаг олон хэмжээст өгөгдлийн статистик шинжилгээний нэг чиглэл юм.

Цэгүүдийн адил төстэй буюу ойр хол байдал нь зай гэдэг ойлголттой холбогдоно. Иймээс тус шинжилгээг цэгүүд хоорондын зайг илэрхийлэх матриц дээр тулгуурлаж хийдэг.

Жишээ 36. www.kaggle.com/unsdsn/world-happiness веб хуудас дээр буй улс орнуудын аз жаргалын индекс, түүнийг тооцож гаргахад ашигласан нэг хүнд ногдох дотоодын нийт бүтээгдэхүүн, нийгмийн халамж, эрүүл мэнд, эрх чөлөө, өгөөмөр байдал, авилга зэрэг хувьсагчдын утгуудыг агуулсан өгөгдөл авч үзье. Тэгвэл улс орнуудыг дээрх 6 хувьсагчийн утгуудаар ойролцоо эсвэл ялгаатай

#	Улс	Оноо	1 хүнд ногдох ДНБ	⋯
1	Финланд	7.769	1.340	⋯
82	Грек	5.287	1.181	⋯
83	Монгол	5.285	0.948	⋯
156	Өмнөд Судан	2.853	0.306	⋯

Хүснэгт 7: Аз жаргалын индекс, 2019 он, өгөгдлийн зарим хэсэг

байдлыг харуулсан цэгэн диаграмм байгуул.

Шинжилгээний гол санаа

Жишээний хувьд түүврийн корреляцийн матриц

$$R = \begin{pmatrix} 1.00 & 0.75 & 0.84 & 0.38 & -0.08 & 0.30 \\ 0.75 & 1.00 & 0.72 & 0.45 & -0.05 & 0.18 \\ 0.84 & 0.72 & 1.00 & 0.39 & -0.03 & 0.30 \\ 0.38 & 0.45 & 0.39 & 1.00 & 0.27 & 0.44 \\ -0.08 & -0.05 & -0.03 & 0.27 & 1.00 & 0.33 \\ 0.30 & 0.18 & 0.30 & 0.44 & 0.33 & 1.00 \end{pmatrix}$$

байгаа нь хувьсагчид өөр хоорондоо хамааралтайг илтгэнэ. Иймд гол хэсгийн шинжилгээ хийж энэхүү холбоо хамаарлыг хамгийн сайн илэрхийлж чадах

[13]Principal Coordinate Analysis эсвэл Multidimensional Scaling

эхний хоёр гол хэсгээр улс орнуудын аз жаргалын ойролцоо эсвэл ялгаатай байдлыг харуулсан цэгэн диаграмм байгуулж болох юм. Гэхдээ тус шинжилгээг хувьсагч буюу баганын хувьд бус харин түүврийн элемент буюу мөрийн хувьд хийнэ.

Өгөгдлийн матрицыг X гэе. Мөн өгөгдөл дэх хувьсагч бүрийн дундаж тэгтэй тэнцүү байг. Тэгвэл түүврийн ковариацийн матриц ба өгөгдлийн матриц хоёр дараах байдлаар холбогддог.

$$S = \frac{1}{n-1} X^T X$$

Цаашилбал саяхан дурдсанчлан гол хэсгийн шинжилгээг баганын хувьд бус харин мөрийн хувьд хийх тул түүврийн ковариацийн матриц буюу угтаа $X^T X$ матрицын хувийн утгын задаргаа бус харин XX^T матрицын хувийн утгын задаргааг ашиглана.

R програмын `cmdscale()` функц

Тус шинжилгээг R програмын тусламжтай хийхэд `cmdscale()` функцийг нь ашигладаг. Үүний тулд тус функцийн `d` аргументаар цэгүүдийн хоорондох зайг илэрхийлсэн матриц дамжуулж өгнө. Зайн матрицыг D гэж тэмдэглэе. Энэ нь

$$D = \left(d_{ij}\right)_{i,j=1\ldots,n}$$

буюу дурын i болон j дүгээр цэгүүдийн хоорондох зайг агуулсан матриц байна. Хэрэв i болон j дүгээр цэгүүдийн хоорондох зайг илэрхийлэхдээ Евклидийн зай ашиглавал $d_{ij}^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$ болно. Зайн матрицыг R програм дээр дараах байдлаар олно.

```
D <- dist(x = X, method = "euclidean")
```

Энд `X` нь өгөгдөл агуулсан датафрейм, тиббл, матриц, хоёр хэмжээст массив зэрэг объект байж болно. Улмаар уг шинжилгээг дараах байдлаар хийнэ.

```
cmdscale(d = D, k = 2)
```

Энд `k` нь огторгуйн хэмжээс заах аргумент юм.

2 Цэгүүдийн хоорондох зай дээр үндэслэн тэдгээрийн координатыг сэргээн олох

Өгөгдлийн матриц

Шинжилгээг зайн матрицад үндэслэж хийх тул эхлээд зайн матрицыг ямар нэг байдлаар өгөгдлийн матрицтай холбох хэрэгтэй.

p огторгуйн хэмжээс буюу санамсаргүй хувьсагчдын тоо

n цэгийн тоо буюу түүврийн хэмжээ

x_i i дүгээр цэгийн координат буюу түүврийн i дүгээр элемент

$$x_i = (x_{i1}, \ldots, x_{ip})^T$$

Энэ нь өгөгдлийн матрицын мөр болно.

X өгөгдлийн матриц

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}^T = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Өгөгдлийн матрицыг зайн матрицтай холбох

Өмнө дурдсан XX^T матрицыг B гэж тэмдэглэе. Өөрөөр хэлбэл

$$B = XX^T$$

гэе. Тэгэхээр $B = \left(b_{ij}\right)_{i,j=1,\dots,n}$ матрицын i дүгээр мөр j дүгээр баганын элемент

$$b_{ij} = x_i^T x_j = \sum_{k=1}^{p} x_{ik} x_{jk}$$

байна. Иймхүү X өгөгдлийн матрицыг D зайн матрицтай холбох нь тус матрицын b_{ij} элементүүдийг зайн матрицын d_{ij} элементүүдээр илэрхийлэх явдал болно.

Зайн матрицыг өгөгдлийн матрицтай холбох

Цэгүүдийн хоорондох зайг хэмжихэд Евклидийн норм ашигласан гэвэл $D = \left(d_{ij}\right)_{i,j=1\dots,n}$ зайн матрицыг $B = \left(b_{ij}\right)_{i,j=1\dots,n} = \left(x_i^T x_j\right)_{i,j=1\dots,n}$ буюу угтаа X өгөгдлийн матрицтай дараах байдлаар холбож болно.

$$\begin{aligned} d_{ij}^2 &= (x_{i1} - x_{j1})^2 + \dots + (x_{ip} - x_{jp})^2 \\ &= x_{i1}^2 + x_{j1}^2 - 2x_{i1}x_{j1} + \dots + x_{ip}^2 + x_{jp}^2 - 2x_{ip}x_{jp} \\ &= x_{i1}^2 + \dots + x_{ip}^2 + x_{j1}^2 + \dots + x_{jp}^2 - 2(x_{i1}x_{j1} + \dots + x_{ip}x_{jp}) \\ &= x_i^T x_i + x_j^T x_j - 2x_i^T x_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned}$$

B матрицын чанар

Чанар 14.

$$\sum_{i=1}^{n} b_{ij} = 0 \quad \forall j = 1, \dots, n$$

Баталгаа Өгөгдлийн матриц дээр хувьсагчдын дундаж тэгтэй тэнцүү буюу $\sum_{k=1}^{p} x_{ik} = 0$ нөхцөл тавьсан тул

$$\sum_{i=1}^{n} b_{ij} = \sum_{i=1}^{n} x_i^T x_j = (x_1 + \dots + x_n)^T x_j = (x_{1k} + \dots + x_{nk})_{k=1,\dots,p}^T x_j = 0$$

болно. □

Зайн матрицаар B матрицыг илэрхийлэх

Евклидийн зайн квадратыг B матрицын элементүүдээр $d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$ гэж илэрхийлсэн. Одоо өмнөх чанарыг тооцон i, j бас i ба j индексүүдээр нийлбэрчилбэл

$$\begin{aligned} d_{\cdot j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 &= \frac{1}{n}\sum_{i=1}^{n} b_{ii} + b_{jj} \\ d_{i\cdot}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 &= b_{ii} + \frac{1}{n}\sum_{j=1}^{n} b_{jj} \\ d_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 &= \frac{2}{n}\sum_{i=1}^{n} b_{ii} \end{aligned}$$

тэгшитгэл гарна. Энд $\cdot$ нь дунджийг илэрхийлнэ. Дээрх дөрвөн тэгшитгэлээс

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2)$$

шийд олдоно.

B матрицын бусад чанар

Чанар 15. 1. B нь тэгш хэмт матриц юм.

2. $\mathrm{rank}(B) = p$

3. B нь эерэг хагас тодорхойлогдсон матриц байна.

Баталгаа 1. $B^T = (XX^T)^T = (X^T)^T X^T = XX^T = B$

2. Хувьсагчдыг шугаман хамааралгүй, тэг дээр бөхсөн тархалттай биш бас цэгүүдийн дор хаяж p ширхэг нь шугаман хамааралгүй гэж тооцсон үед $\mathrm{rank}(B) = \mathrm{rank}(XX^T) = \mathrm{rank}(X) = p$ болно.

3. Зайг хэмжихдээ Евклидийн норм ашигласан гэдгээс мөрдөн гарна.

□

B матрицын хувийн утгын задаргаа, цэгүүдийн координат

Мөрдлөгөө 2. B матриц p ширхэг эерэг, $n - p$ ширхэг тэгтэй тэнцүү хувийн утгатай.

$$B = \Gamma\Lambda\Gamma^T$$

Энд $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ нь B матрицын хувийн утгуудаас тогтох матриц, $\Gamma = (\gamma_1, \ldots, \gamma_p)$ нь хувийн векторуудаас тогтох матриц юм. $B = XX^T$ гэдгийг санавал

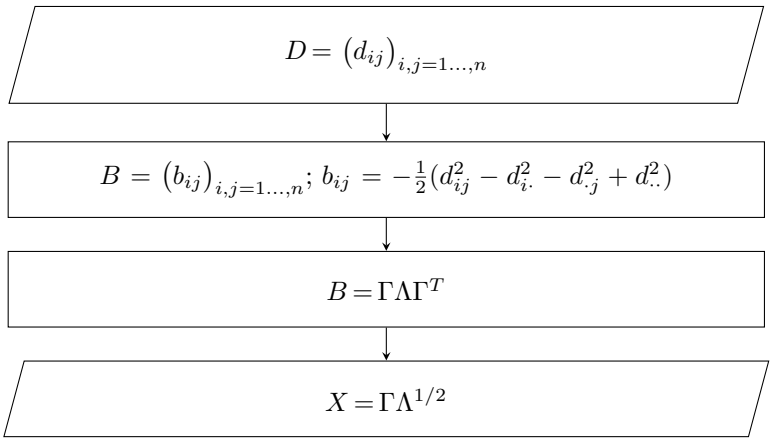
$$B = \Gamma\Lambda\Gamma^T = \Gamma\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\Gamma^T = \underbrace{\Gamma\Lambda^{\frac{1}{2}}}_{X}\underbrace{\Lambda^{\frac{1}{2}}\Gamma^T}_{X^T} = XX^T$$

болно. Энэ нь өгөгдлийн матриц буюу цэгүүдийн координат

$$X = \Gamma\Lambda^{\frac{1}{2}}$$

гэж тодорхойлогдоно гэсэн үг юм.

Цэгүүдийн координат сэргээх үйлдлийн дэс дараа



Зураг 24: Цэгүүдийн координатыг тэдгээрийн хоорондох зай дээр үндэслэн сэргээн олох үйлдлийн дэс дараа

3 Огторгуйн хэмжээс сонгох

Огторгуйн хэмжээс сонгох

Шинжилгээний үндсэн зорилго нь их хэмжээст огторгуй дахь цэгүүдийн адил төстэй байдлыг бага хэмжээст огторгуйд тухайлбал хавтгайд буулгаж дүрслэх явдал тул огторгуйн хэмжээсийг ихэвчлэн $p = 2$ гэж авдаг. Огторгуйн хэмжээсийг B матрицын ранг эсвэл тэг биш хувийн утгуудын тоогоор сонгож болох хэдий ч практикт хэмжээс бууруулах зорилго их тулгардаг тул хувийн утгуудын харьцаанд үндэслэж хэмжээс сонгодог.

$$\psi_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \mathrm{abs}(\lambda_i)} \quad \text{эсвэл} \quad \psi_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \max(\lambda_i, 0)}$$

Энд $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ гэж тооцов.

Жишээний хувьд $p = 2$ үед

$$\psi_2 \approx \frac{41.845 + 4.541}{53.536} \approx 0.866$$

буюу нийт ковариацийн ойролцоогоор 86.6 хувь нь 2 хэмжээст шинэ огторгуйд хадгалагдан үлдэж байна.

ψ_p харьцааны хоёр хувилбарын тоон утгыг `cmdscale()` функцээр олох боломжтой. Үүний тулд `eig` аргументаар нь `TRUE` утга дамжуулна. Тэгвэл уг харьцаануудын тоон утгаас гадна $\lambda_1, \ldots, \lambda_n$ хувийн утгуудыг тус функцийн буцаах утгын `GOF` болон `eig` элементүүдийн утга байдлаар олгосон байдаг. Харин энэ тохиолдолд цэгүүдийн проекцын координатыг гарган авахын `points` элементийг дуудах шаардлагатай.

4 Хувьсагчдын масштабын ялгаатай байдлын нөлөөг анхаарах

Масштабын ялгаатай байдлын нөлөөг анхаарах

Хувьсагчдын хэмжээс тухайлбал хэмжүүрийн нэгж ялгаатай үед зарим хувьсагч зайн хэмжээст хэт давамгайлах байдал үүсэх талтай. Улмаар энэ нь тус шинжилгээний үр дүн болох цэгүүдийн хол, ойр байдлыг бодит байдлаас гажуудуулдаг. Хувьсагчдын масштабын нөлөө, түүнийг хэрхэн зайлуулах, хувьсагчдын масштабыг тэгшитгэх талаар кластерын шинжилгээ, гол хэсгийн шинжилгээ зэрэг сэдэвт үзсэн.

Лекц XIV
Конжойнт шинжилгээ

1 Конжойнт шинжилгээ

Конжойнт шинжилгээ

Конжойнт шинжилгээ нь бараа, үйлчилгээний онцлогийг хэрэглэгчид хэрхэн тодорхойлж буйг судлах зорилготой, маркетингийн судалгааны нэгэн статистик арга юм. Уг шинжилгээг

- шинээр гаргах бараа, үйлчилгээнд тусгаж болох онцлог шинжүүдээс чухам аль нь хэрэглэгчийн таашаалд илүү нийцэхийг олох
- бараа, үйлчилгээний олон янзын онцлог шинж бүхий хувилбаруудаас оновчтойг нь сонгох
- бараа, үйлчилгээний аль шинжийг түлхүү сурталчлах буюу зар, сурталчилгааны стретаги тогтоох

зэрэг зорилгод ашигладаг.

Энэхүү шинжилгээг эдийн засаг, маркетингийн талаас нь бус харин цэвэр статистикийн зүгээс авч үзнэ.

Аль хувилбар хамгийн сайн бэ?

Орц	Сав	
	жигнэмэг	хуванцар
аарц	?	?
шоколад	?	?
жимс	?	?

$X_1, \ldots, X_K$ бараа үйлчилгээний онцлогийг тодорхойлох фактор жишээлбэл X_1 нь орц найрлага, X_2 нь сав баглаа боодол гэх мэт

$L_1, \ldots, L_K$ $X_1, \ldots, X_K$ фактор тус бүрийн хувилбарын тоо

$M = L_1 \cdot \ldots \cdot L_K$ бараа үйлчилгээний нийт хувилбарын тоо

Хэрэглэгч таашаалараа хувилбаруудад оноо өгсөн нь

Орц	**Сав**	
	жигнэмэг	хуванцар
аарц	2	1
шоколад	6	3
жимс	5	4

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \\ 5 \\ 1 \\ 3 \\ 4 \end{pmatrix}$$

Хэрэглэгчийн сэтгэл ханамжийг илэрхийлэх дээрх оноонууд бол хувилбаруудын аль нь алинаасаа илүү гэдгийг заахаас хэтрэхгүй. Үнэн хэрэгтээ Y_i нь үүнээс илүү нарийвчлалтай тоон утгатай байна. Иймд $\hat{Y}_i$ буюу тэрхүү жинхэнэ утгын үнэлэлт шаардлагатай юм.

Түүнчлэн уг оноо нь заавал эерэг, бүхэл, дэс дараалсан тоо байх албагүй. Өөрөөр хэлбэл дурын бодит тоо байж болохын зэрэгцээ давтагдаж болно.

R програмын conjoint багц ашиглаж $\hat{Y}_i$ үнэлэлт олох

```
conjoint::caTotalUtilities(
  y = matrix(data = c(
    2,6,5,1,3,4
  ), nrow = 1, byrow = TRUE),
  x = expand.grid(
    ingredients = c("curd", "chocolate", "fruit"),
    packing = c("biscuit", "plastic")
  )
)
```

```
2.333 5.333 5.333 0.667 3.667 3.667
```

$$Y = \begin{pmatrix} 2 \\ 6 \\ 5 \\ 1 \\ 3 \\ 4 \end{pmatrix} \quad \hat{Y} = \begin{pmatrix} 2.333 \\ 5.333 \\ 5.333 \\ 0.667 \\ 3.667 \\ 3.667 \end{pmatrix}$$

2 Конжойнт шинжилгээний загвар

Конжойнт шинжилгээний загвар

Бараа үйлчилгээний i дүгээр хувилбараас авах сэтгэл ханамж нь түүний онцлогийг тодорхойлох фактор тус бүрийн өгөх сэтгэл ханамжаас бүтнэ. Харин тухайн нэг фактор хэр зэрэг сэтгэл ханамж өгөх нь факторын чухам ямар

хувилбарыг тус бараа үйлчилгээнд тусгаснаас шалтгаална.

$$Y_i = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \beta_{kl} \, \mathrm{I}(X_k = x_{kl}) + \mu + \epsilon_i, \quad \forall k : \sum_{l=1}^{L_k} \beta_{kl} = 0$$

X_k фактор $(k = 1, \ldots, K)$

x_{kl} X_k факторын нэг хувилбар $(l = 1, \ldots, L_k)$

β_{kl} X_k факторын x_{kl} хувилбараас авах сэтгэл ханамжийг илэрхийлэх оноо

μ дундаж оноо

ϵ_i загварын алдаа

Ийнхүү $L_1 + \ldots + L_K$ ширхэг үл мэдэгдэх параметр бүхий $M + K$ ширхэг тэгшитгэлтэй шугаман загвар зохиолоо.

Конжойнт шинжилгээний загварын зарим шинж чанар

1. Конжойнт шинжилгээний загвар нь нөхцөлт шугаман регрессийн загвар юм.
2. $\sum\limits_{l=1}^{L_k} \beta_{kl} = 0$ нөхцөл нь бараа үйлчилгээний тухайн нэг хувилбарт факторын бүх хувилбарыг оруулах нь тус факторын ялгарлыг үгүй болгох тул уг факторыг тооцоогүйтэй адил болохыг илэрхийлнэ.

Жишээний хувьд бичигдэх конжойнт шинжилгээний загвар

Жишээний хувьд 6 үл мэдэгдэх параметр бүхий 8 тэгшитгэлээр тодорхойлогдох загвар зохиогдоно.

$$\begin{cases} Y_1 = \beta_{11} + \beta_{21} + \mu \\ Y_2 = \beta_{12} + \beta_{21} + \mu \\ Y_3 = \beta_{13} + \beta_{21} + \mu \\ Y_4 = \beta_{11} + \beta_{22} + \mu \\ Y_5 = \beta_{12} + \beta_{22} + \mu \\ Y_6 = \beta_{13} + \beta_{22} + \mu \\ \beta_{11} + \beta_{12} + \beta_{13} = 0 \\ \beta_{21} + \beta_{22} = 0 \end{cases}$$

$Y_1, \ldots, Y_6$ буюу бараанаас авах сэтгэл ханамжийн жинхэнэ оноог үнэлэхийн тулд эхлээд уг загварын $\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}$ болон μ параметрүүдийг үнэлэх шаардлагатай.

3 Загварын параметрийн үнэлэлт

Загварын параметр үнэлэх хүснэгтийн арга

Загварын параметрийг өгөгдөл агуулсан хүснэгтийн мөр, багана болон нийт утгуудын дундаж улмаар тэдгээрийн ялгаврын тусламжтай хялбар үнэлж болдог. Энэ нь $\hat{\beta}_{11} = -2$, $\hat{\beta}_{12} = 1$, $\hat{\beta}_{13} = 1$, $\hat{\beta}_{21} = 0.833$, $\hat{\beta}_{22} = -0.833$ болон $\hat{\mu} = 3.5$ үнэлэлт олсон гэсэн үг юм.

X_1	X_2 1	2	$\overline{Y}_{X_1\cdot}$	$\hat{\beta}_{1l}$
1	2	1	$\frac{2+1}{2}=1.5$	$1.5-3.5=-2$
2	6	3	4.5	1
3	5	4	4.5	1
$\overline{Y}_{X_2\cdot}$	$\frac{2+6+5}{3}=4.333$	2.666	$\frac{2+1+6+3+5+4}{6}=3.5$	
$\hat{\beta}_{2l}$	$4.333-3.5=0.833$	-0.833		

Бараа үйлчилгээнээс авах сэтгэл ханамжийн жинхэнэ оноог үнэлэх

Загварын параметрүүд $\hat{\beta}_{11}=-2$, $\hat{\beta}_{12}=1$, $\hat{\beta}_{13}=1$, $\hat{\beta}_{21}=0.833$, $\hat{\beta}_{22}=-0.833$ ба $\hat{\mu}=3.5$ гэж олдсон тул бараанаас авах сэтгэл ханамжийн онооны жинхэнэ утгыг загварын дагуу дараах байдлаар үнэлнэ.

$$\hat{Y}_1=\hat{\beta}_{11}+\hat{\beta}_{21}+\hat{\mu}=-2+0.833+3.5=2.333$$
$$\hat{Y}_2=\hat{\beta}_{12}+\hat{\beta}_{21}+\hat{\mu}=1+0.833+3.5=5.333$$
$$\hat{Y}_3=\hat{\beta}_{13}+\hat{\beta}_{21}+\hat{\mu}=1+0.833+3.5=5.333$$
$$\hat{Y}_4=\hat{\beta}_{11}+\hat{\beta}_{22}+\hat{\mu}=-2-0.833+3.5=0.667$$
$$\hat{Y}_5=\hat{\beta}_{12}+\hat{\beta}_{22}+\hat{\mu}=1-0.833+3.5=3.667$$
$$\hat{Y}_6=\hat{\beta}_{13}+\hat{\beta}_{22}+\hat{\mu}=1-0.833+3.5=3.667$$

Энэ нь `conjoint` багцын `caTotalUtilities()` функцээр олсонтой тохирно.

Олон хэрэглэгчийн сэтгэл ханамжийн судалгааны өгөгдлөөр конжойнт шинжилгээний загварын параметр үнэлэх

X_1	X_2 1	2	3
1	1	3	4
2	2	5	6

хэрэглэгч Ц

X_1	X_2 1	2	3
1	2	4	3
2	1	5	6

хэрэглэгч Д

X_1	X_2 1	2	3	$\overline{Y}_{X_1\cdot}$	$\hat{\beta}_{1l}$
1	1,2	3,4	4,3	$\frac{1+2+3+4+4+3}{6}=2.833$	$2.833-3.5=-0.666$
2	2,1	5,5	6,6	4.166	0.666
$\overline{Y}_{X_2\cdot}$	1.5	4.25	4.75	3.5	
$\hat{\beta}_{2l}$	-2	0.75	1.25		

R програмын `conjoint` багц ашиглаж загварын параметр үнэлэх

```
conjoint::caUtilities(
  y = matrix(data = c(
```

```
    1,2,3,5,4,6,
    2,1,4,5,3,6
  ), nrow = 2, byrow = TRUE),
  x = expand.grid(
    X1 = c("X11", "X12"),
    X2 = c("X21", "X22", "X23")
  ),
  z = c("X11", "X12", "X21", "X22", "X23")
)
```

```
3.5000000 -0.6666667  0.6666667 -2.0000000  0.7500000  1.2500000
```

Хэрэглэгч Ц ба Д нарын бараа үйлчилгээнээс авах сэтгэл ханамжийн дундаж онооны үнэлэлт

Загварын параметрийг олон хэрэглэгчээс авсан судалгааны өгөгдлөөр үнэлсэн тул мөнөөх $\hat{\beta}_{11} = -0.666$, $\hat{\beta}_{12} = 0.666$, $\hat{\beta}_{21} = -2$, $\hat{\beta}_{22} = 0.75$, $\hat{\beta}_{23} = 1.25$ ба $\hat{\mu} = 3.5$ утгуудыг орлуулан бодож олсон бараа үйлчилгээнээс авах сэтгэл ханамжийн оноо нь хэрэглэгчдийн сэтгэл ханамжийн дундаж утгыг үнэлнэ.

$$\hat{Y}_1 = \hat{\beta}_{11} + \hat{\beta}_{21} + \hat{\mu} = -0.666 - 2 + 3.5 = 0.834$$
$$\hat{Y}_2 = \hat{\beta}_{12} + \hat{\beta}_{21} + \hat{\mu} = 0.666 - 2 + 3.5 = 2.166$$
$$\hat{Y}_3 = \hat{\beta}_{11} + \hat{\beta}_{22} + \hat{\mu} = -0.666 + 0.75 + 3.5 = 3.584$$
$$\hat{Y}_4 = \hat{\beta}_{12} + \hat{\beta}_{22} + \hat{\mu} = 0.666 + 0.75 + 3.5 = 4.916$$
$$\hat{Y}_5 = \hat{\beta}_{11} + \hat{\beta}_{23} + \hat{\mu} = -0.666 + 1.25 + 3.5 = 4.084$$
$$\hat{Y}_6 = \hat{\beta}_{12} + \hat{\beta}_{23} + \hat{\mu} = 0.666 + 1.25 + 3.5 = 5.416$$

R програмын conjoint багц ашиглаж хэрэглэгч тус бүрийн болон нийт хэрэглэгчдийн сэтгэл ханамжийн дундаж утгыг үнэлэх

```
conjoint::caTotalUtilities( # хэрэглэгч нэг бүрийн сэтгэл ханамж
  y = c(
    1,2,3,5,4,6,
    2,1,4,5,3,6
  ),
  x = expand.grid(
    X1 = c("X11", "X12"),
    X2 = c("X21", "X22", "X23")
  )
) |> colMeans() # хэрэглэгчдийн сэтгэл ханамжийн дундаж утга
```

```
0.8335 2.1665 3.5835 4.9165 4.0835 5.4165
```

4 Конжойнт шинжилгээний загварыг энгийн шугаман регрессийн загвар руу хувиргах нь

Загварыг энгийн шугаман регресс рүү хувиргах нь

Конжойнт шинжилгээний өргөтгөсөн шугаман регрессийн загварыг нөхцөлгүй буюу

$$Y = X\beta$$

хэлбэртэй энгийн шугаман регрессийн загварт хувиргах боломжтой. Үүний тулд эхлээд шинээр зохиох загвар дахь үл мэдэгдэх параметрийн тоог тогтоох шаардлагатай. $\sum_{l=1}^{L_k} \beta_{kl} = 0$ $(k = 1, \ldots, K)$ нөхцлүүдийг тооцвол үл мэдэгдэгчдийн тоо K-аар цөөрнө. Иймд нийт үл мэдэгдэгчдийн тоо $N = \sum_{k=1}^{K} L_k - K + 1$ болох тул $\beta = (\beta_1, \ldots, \beta_N)^T$ болно. Энд хамгийн сүүлд нь μ дунджийг тооцож нэгийг нэмэв.

Зайрмагны жишээний хувьд бичигдэх энгийн шугаман регрессийн загварын параметрийн тоо $N = (3 + 2) - 2 + 1 = 4$ байна.

Энгийн шугаман регрессийн загвар зохиох тухай

$Y = X\beta$ энгийн шугаман регрессийн загвар зохиохдоо β параметрийг хүсэнээрээ сонгох боломжтой. Харин X матриц ямар байх нь β параметрийн сонголтоос хамаардаг. X нь β параметр болон анхны загвар хоёроос хамаарч зохиогдох дамми хувьсагчдаас тогтсон, M (нийт хувилбарын тоо) мөр ба N (үл мэдэгдэгчдийн тоо) баганатай матриц байна.

β параметрийг зохиох байдал

Зайрмагны жишээний хувьд

$$\begin{aligned} Y_1 &= \beta_{11} + \beta_{21} + \mu & \qquad Y_4 &= \beta_{11} + \beta_{22} + \mu \\ Y_2 &= \beta_{12} + \beta_{21} + \mu & \qquad Y_5 &= \beta_{12} + \beta_{22} + \mu \\ Y_3 &= \beta_{13} + \beta_{21} + \mu & \qquad Y_6 &= \beta_{13} + \beta_{22} + \mu \end{aligned}$$

байсныг анхаарааД $\beta_1 = Y_6$ гэж эхэлбэл β параметр дараах байдлаар зохиогдоно.

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} Y_6 \\ Y_5 - Y_6 \\ Y_4 - Y_6 \\ Y_3 - Y_6 \end{pmatrix} = \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ (\beta_{12} + \beta_{22} + \mu) - (\beta_{13} + \beta_{22} + \mu) \\ (\beta_{11} + \beta_{22} + \mu) - (\beta_{13} + \beta_{22} + \mu) \\ (\beta_{13} + \beta_{21} + \mu) - (\beta_{13} + \beta_{22} + \mu) \end{pmatrix} = \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ \beta_{12} - \beta_{13} \\ \beta_{11} - \beta_{13} \\ \beta_{21} - \beta_{22} \end{pmatrix}$$

X матрицыг олох байдал

$Y = X\beta$ буюу

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} \beta_{11} + \beta_{21} + \mu \\ \beta_{12} + \beta_{21} + \mu \\ \beta_{13} + \beta_{21} + \mu \\ \beta_{11} + \beta_{22} + \mu \\ \beta_{12} + \beta_{22} + \mu \\ \beta_{13} + \beta_{22} + \mu \end{pmatrix} = \begin{pmatrix} ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix} \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ \beta_{12} - \beta_{13} \\ \beta_{11} - \beta_{13} \\ \beta_{21} - \beta_{22} \end{pmatrix}$$

гэдгийг анхаарна. Эндээс дараах матриц олдоно.

$$X = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$Y = X\beta$ загварын β параметрийн үнэлгээ

Загварыг хамгийн бага квадратын аргаар үнэлнэ. Бидний зохиосон β параметрийн хувьд β_1 параметрт харгалзах X матрицын нэгдүгээр багана тогтмол буюу дан нэгээс тогтсон учраас уг параметр загварын сул гишүүнээр үнэлэгдэнэ. Өөрөөр хэлбэл X матрицын нэг дүгээрхээс бусад баганыг тайлбарлах хувьсагч болгож авна. Тодруулбал R програмын хувьд тус загварыг дараах байдлаар томьёолж бичнэ.

```
fit <- lm(formula = Y ~ X[,-1])
```

Жишээний хувьд түүврийн хэмжээ $n = 1$ байх эхний тохиолдолд тус загварын параметрүүд

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{pmatrix} = \begin{pmatrix} 3.(6) \\ 0 \\ -3 \\ 1.(6) \end{pmatrix}$$

гэж олдоно.

Конжойнт шинжилгээний загварын β_{kl} параметрүүдийг олох

$$\begin{cases} \hat{\beta} = \begin{pmatrix} 3.(6) \\ 0 \\ -3 \\ 1.(6) \end{pmatrix} = \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ \beta_{12} - \beta_{13} \\ \beta_{11} - \beta_{13} \\ \beta_{21} - \beta_{22} \end{pmatrix} \\ \begin{cases} \beta_{11} + \beta_{12} + \beta_{13} = 0 \\ \beta_{21} + \beta_{22} = 0 \end{cases} \end{cases} \qquad \begin{pmatrix} \hat{\beta}_{11} \\ \hat{\beta}_{12} \\ \hat{\beta}_{13} \\ \hat{\beta}_{21} \\ \hat{\beta}_{22} \\ \hat{\mu} \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0.8(3) \\ -0.8(3) \\ 3.5 \end{pmatrix}$$

```
conjoint::caPartUtilities(
  y = matrix(data = c(2,6,5,1,3,4), nrow = 1, byrow = TRUE),
```

```
x = expand.grid(ingredients = c("curd", "chocolate", "fruit"),
  packing = c("biscuit", "plastic")),
z = c("curd", "chocolate", "fruit", "biscuit", "plastic"))
```

```
intercept curd chocolate fruit biscuit plastic
      3.5   -2         1     1   0.833  -0.833
```

Олон хэрэглэгчээс цуглуулсан өгөгдлийн хувьд параметр үнэлэх

Түүврийн нэг элементийг $Y^{(i)}$ гэвэл шинээр зохиосон $Y = X\beta$ нөхцөлт бус шугаман регрессийн загварыг

$$Y^* = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(n)} \end{pmatrix} \qquad X^* = \left.\begin{pmatrix} X \\ X \\ \vdots \\ X \end{pmatrix}\right\} n \text{ удаа}$$

байх

$$Y^* = X^*\beta$$

хэлбэрт шилжүүлээд β параметрийг үнэлнэ.

Жишээ болгон $n = 3$ хэмжээтэй дараах өгөгдөл авъя. Тэгвэл өмнөх слайд

	X_2	
	3	1
X_1	5	2
	6	4

	X_2	
	1	2
X_1	4	3
	6	5

	X_2	
	1	3
X_1	4	2
	5	6

дээрх аргачлалын дагуу дараах үр дүнд хүрнэ.

$$\hat\beta = \begin{pmatrix} 4.9(4) \\ -2 \\ -3.5 \\ 0.(7) \end{pmatrix} \qquad \begin{pmatrix} \hat\beta_{11} \\ \hat\beta_{12} \\ \hat\beta_{13} \\ \hat\beta_{21} \\ \hat\beta_{22} \\ \hat\mu \end{pmatrix} = \begin{pmatrix} -1.(6) \\ -0.1(6) \\ 1.8(3) \\ 0.3(8) \\ -0.3(8) \\ 3.5 \end{pmatrix} \qquad \widehat{\overline{Y}} = \begin{pmatrix} 2.(2) \\ 3.7(2) \\ 5.7(2) \\ 1.(4) \\ 2.9(4) \\ 4.9(4) \end{pmatrix}$$

Энэ нь хүснэгтийн арга болон `conjoint` багцын тусламжтай олсонтой ижил байна.

Лекц XV
Каноник корреляцын шинжилгээ

1 Каноник корреляц

Каноник корреляцын шинжилгээ

Каноник корреляцаар хоёр санамсаргүй векторын холбоо хамаарлыг хэмждэг.

№	X_1	X_2	X_3	Y_1	Y_2	№	X_1	X_2	X_3	Y_1	Y_2
1	62	64	60	75	60	7	63	72	60	72	76
2	82	94	65	90	85	8	70	80	60	69	60
3	78	82	80	68	71	9	65	61	70	61	80
4	60	63	82	71	80	10	65	62	63	66	70
5	60	61	70	60	60	11	89	97	90	93	95
6	60	65	63	72	70	12	65	60	66	61	85

Хүснэгт 10: Статистикийн хөтөлбөрөөр суралцаж төгссөн 12 оюутны дүн; X_1 математик анализ, X_2 шугаман алгебр, аналитик геометр, X_3 дискрет математик, математик логик, Y_1 магадлалын онол, Y_2 математик статистик

Жишээ 37. Хичээлүүдийн залгамж холбоог каноник корреляцаар шинжил.

R програмын CCA багцын `cc()` функц

Өгөгдөл оруулах байдал

```
X <- matrix(
  data = c(62, 82, 78, 60, 60, 60, 63, 70, 65, 65, 89, 65, 64,
    94, 82, 63, 61, 65, 72, 80, 61, 62, 97, 60, 60, 65, 80, 82,
    70, 63, 60, 60, 70, 63, 90, 66),
  ncol = 3, byrow = FALSE)
Y <- matrix(
  data = c(75, 90, 68, 71, 60, 72, 72, 69, 61, 66, 93, 61, 60,
    85, 71, 80, 60, 70, 76, 60, 80, 70, 95, 85),
  ncol = 2, byrow = FALSE)
```

Шинжилгээ хийх байдал ба каноник корреляц

```
cca <- CCA::cc(X, Y)
cca$cor
```

Үр дүн

```
0.8453653 0.6224044
```

Каноник корреляцын коэффициент

$$a^T X = a_1 X_1 + \ldots + a_q X_q$$

болон

$$b^T Y = b_1 Y_1 + \ldots + b_p Y_p$$

шугаман эвлүүлгүүдийн корреляцын хамгийн их

$$\max_{a,b} \operatorname{cor}(a^T X, b^T Y)$$

утгыг *каноник корреляцын коэффициент* гэнэ.

2 Шугаман эвлүүлгүүдийн корреляц

a^TX ба b^TY хувьсагчдын корреляцын коэффициент

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \right)$$

гэе. Энд

$$\begin{aligned} \Sigma_{XX} &= \operatorname{cov}(X) && (q \times q) \\ \Sigma_{YY} &= \operatorname{cov}(Y) && (p \times p) \\ \Sigma_{XY} &= \Sigma_{YX}^T = \operatorname{cov}(X,Y) = E(X-\mu)(Y-\nu)^T && (q \times p) \end{aligned}$$

байна. Тэгвэл a^TX ба b^TY скаляр санамсаргүй хувьсагчдын корреляцын коэффициентыг

$$\operatorname{cor}(a^TX, b^TY) = \frac{\operatorname{cov}(a^TX, b^TY)}{(\operatorname{cov}(a^TX))^{1/2}(\operatorname{cov}(b^TY))^{1/2}} = \frac{a^T\Sigma_{XY}b}{(a^T\Sigma_{XX}a)^{1/2}(b^T\Sigma_{YY}b)^{1/2}}$$

байдлаар олж болно.

Каноник корреляцын бодлого

$\forall c \in \mathbb{R}^+$ тогтмолын хувьд $\operatorname{cor}(c \cdot \xi, \eta) = \operatorname{cor}(\xi, \eta)$ байдаг өөрөөр хэлбэл корреляцын коэффициент масштабаас хамаардаггүй тул

$$\begin{aligned} \operatorname{cov}(a^TX) &= a^T\Sigma_{XX}a = 1 \\ \operatorname{cov}(b^TY) &= b^T\Sigma_{YY}b = 1 \end{aligned}$$

буюу a^TX ба b^TY хувьсагчдын дисперсийг нэгтэй тэнцүү гэж тооцох боломжтой. Ингэснээр $\operatorname{cor}(a^TX, b^TY)$ корреляцын коэффициент нь ердөө $a^T\Sigma_{XY}b$ гэсэн энгийн илэрхийлэлтэй болно. Иймхүү каноник корреляцын коэффициент олохын тулд

$$\begin{aligned} \max_{a,b} \quad & a^T\Sigma_{XY}b \\ \text{s.t.} \quad & a^T\Sigma_{XX}a = 1, \\ & b^T\Sigma_{YY}b = 1 \end{aligned}$$

зааглалттай оптимизацын бодлого бодох боллоо.

Нөгөө талаас санамсаргүй векторууд тус бүр дэх хувьсагчдыг хамааралгүй бас дисперсийг нь нэгтэй тэнцүү болгох буюу Махаланобис хувиргалт шиг $\Sigma_{XX}^{-1/2}X$, $\Sigma_{YY}^{-1/2}Y$ байдлаар хувиргана гэвэл шугаман эвлүүлгийн векторуудыг үүний урвуугаар $\Sigma_{XX}^{1/2}a$, $\Sigma_{YY}^{1/2}b$ гэж хувиргах хэрэгтэй болно. Нэгэнт $\Sigma_{XX}^{-1/2}X$ болон $\Sigma_{YY}^{-1/2}Y$ вектор дахь хувьсагчид дотооддоо хамааралгүй болох тул X болон Y дэх хувьсагчдын дотоод холбоо хамаарлын мэдээлэл харгалзан $\Sigma_{XX}^{1/2}a$ болон $\Sigma_{YY}^{1/2}b$ векторт шилжих юм. Тэгвэл уг хувиргалтаар үүсэх санамсаргүй векторуудын ковариац

$$\operatorname{cov}(\Sigma_{XX}^{-1/2}X, \Sigma_{YY}^{-1/2}Y) = \Sigma_{XX}^{-1/2}\operatorname{cov}(X,Y)\Sigma_{YY}^{-1/2} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$$

болно. Тэгэхээр уг оптимизацын бодлогын зорилгын функцийг

$$a^T\Sigma_{XY}b = a^T\Sigma_{XX}^{1/2}\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\Sigma_{YY}^{1/2}b$$

хэлбэрээр харах хэрэгтэй.

Зорилгын функц дэх ковариацийн матрицын сингуляр утгын задаргаа

$$a^T\Sigma_{XY}b = a^T\Sigma_{XX}^{1/2}\underbrace{\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}}_{K}\Sigma_{YY}^{1/2}b$$

гээд улмаар K матрицын $K = \Gamma\Lambda\Delta^T$ сингуляр утгын задаргаа оруулж ирье. Энд

- $\Gamma = (\gamma_1, \ldots, \gamma_k)$, $\Delta = (\delta_1, \ldots, \delta_k)$, $\Lambda = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})$
- $k = \mathrm{rank}(K) \leq \min\{q, p\}$
- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ нь KK^T эсвэл K^TK матрицуудын тэг биш хувийн утгууд
- γ_i ба δ_j нь харгалзан KK^T ба K^TK матрицуудын хувийн векторууд

байна.

Хэрэв зааглалтын нөхцөл хангах

$$a = \Sigma_{XX}^{-1/2}\gamma_i$$
$$b = \Sigma_{YY}^{-1/2}\delta_i$$

орлуулга хийвэл зорилгын функцийн утга

$$a^T\Sigma_{XY}b = a^T\Sigma_{XX}^{1/2}\Gamma\Lambda\Delta^T\Sigma_{YY}^{1/2}b = \gamma_i^T\Sigma_{XX}^{-1/2}\Sigma_{XX}^{1/2}\Gamma\Lambda\Delta^T\Sigma_{YY}^{1/2}\Sigma_{YY}^{-1/2}\delta_i = \gamma_i^T\Gamma\Lambda\Delta^T\delta_i = \lambda_i^{1/2}$$

болно. Энэхүү шийд ямар нөхцөлд хүчинтэй байхыг тодруулъя. $\gamma_i = \Sigma_{XX}^{1/2}a$ ба $\delta_i = \Sigma_{YY}^{1/2}b$ бас хувийн векторууд ортогонал чанартай тул $i \neq j$ үед

$$\gamma_i^T\gamma_j = a_i^T\Sigma_{XX}^{1/2}\Sigma_{XX}^{1/2}a_j = a_i^T\Sigma_{XX}a_j = 0$$
$$\delta_i^T\delta_j = b_i\Sigma_{YY}^{1/2}\Sigma_{YY}^{1/2}b_j = b_i\Sigma_{YY}b_j = 0$$

байна. Мөн бодлогыг анх томьёолохдоо тавьсан зааглалт ч энд хамаарна.

Бодлогын шийд

Олсон шийдээ теорем болгон томьёолж бичив.

Теорем 10. Өгсөн r $(1 \leq r \leq k)$ бүрийн хувьд

$$a^T\Sigma_{XX}a = 1, \quad b^T\Sigma_{YY}b = 1$$

болон

$$a_i^T\Sigma_{XX}a = 0, \quad b_i^T\Sigma_{YY}b = 0, \quad i \neq r$$

нөхцөлд

$$\max_{a,b} a^T\Sigma_{XY}b$$

хэмжигдэхүүн хамгийн их $\lambda_r^{1/2}$ утгадаа $a = a_r$ болон $b = b_r$ үед хүрнэ.

Бодолтын үр дүн

Шугаман эвлүүлгийн буюу проекцын вектор

$$a_i = \Sigma_{XX}^{-1/2}\gamma_i \qquad b_i = \Sigma_{YY}^{-1/2}\delta_i$$

Эдгээр нь `CCA` багцын `cc()` функцийн буцаах утгын `xcoef` болон `ycoef` элементүүдэд агуулагдана.

Шугаман эвлүүлгээр үүсэх хувьсагч

$$\eta_i = a_i^T X = \gamma_i^T \Sigma_{XX}^{-1/2} X \qquad \varphi_i = b_i^T Y = \delta_i^T \Sigma_{YY}^{-1/2} Y$$

Эдгээрийг *каноник хувьсагч* гэнэ. Хэрэв $EX = 0$ ба $EY = 0$ нөхцөл тавьсан бол эндээс олдох каноник хувьсагчийн утга `CCA::cc()` функцийнхтэй адил болно. Каноник хувьсагчид нь тус функцийн буцаах утгын `scores` элементийн `xscores` болон `yscores` дэд элементүүдэд агуулагддаг.

Каноник корреляцын коэффициент

$$\rho_i = \operatorname{cor}(\eta_i, \varphi_i) = \operatorname{cor}(a_i^T X, b_i^T Y) = \lambda_i^{1/2}$$

Каноник корреляцын чанар

Чанар 16. 1. Каноник хувьсагчдын ковариац дараах томьёоллоор илэрхийлэгдэх бүтэцтэй байдаг.

$$\operatorname{cov}(\eta) = I_k \qquad \operatorname{cov}(\varphi) = I_k \qquad \operatorname{cov}(\eta, \varphi) = \Lambda$$

Энд $\eta = (\eta_1, \ldots, \eta_k)$ ба $\varphi = (\varphi_1, \ldots, \varphi_k)$ бас I_k нь k-хэмжээст нэгж матриц юм.

2. Каноник корреляцын коэффициент шугаман хувиргалтаар инвариант чанартай.

Баталгаа 1.

$$\begin{aligned}\operatorname{cov}(\eta_i, \eta_j) &= \operatorname{cov}(a_i^T X, a_j^T X) = a_i^T \operatorname{cov}(X, X) a_j = a_i^T \Sigma_{XX} a_j \\ &= \gamma_i^T \Sigma_{XX}^{-1/2} \Sigma_{XX} \Sigma_{XX}^{-1/2} \gamma_j = \gamma_i^T \gamma_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}\end{aligned}$$

$\operatorname{cov}(\varphi_i, \varphi_j)$ ковариацад харгалзах тооцоо үүнтэй төстэй.

$$\begin{aligned}\operatorname{cov}(\eta_i, \varphi_j) &= \operatorname{cov}(a_i^T X, b_j^T Y) = a_i^T \operatorname{cov}(X, Y) b_j = a_i^T \Sigma_{XY} b_j \\ &= \gamma_i^T \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} \delta_j = \gamma_i^T K \delta_j = \gamma_i^T \Gamma \Lambda \Delta^T \delta_j = \lambda_{ij}\end{aligned}$$

Энд λ_{ij} нь Λ матрицын i дүгээр мөр, j дүгээр баганын элемент юм. Одоо дээрх үр дүнгүүдийг матриц хэлбэрээр томьёолбол 1 дүгээр чанар гарна.

2. Каноник корреляц нь Пирсоны корреляцын коэффициентоор тодорхойлогдох бөгөөд тэр нь масштаб болон параллель зөөлтөөс болж өөрчлөгддөггүй тул 2 дугаар чанар илэрхий юм.

□

3 Каноник корреляцын тухай таамаглал

Каноник корреляцын тухай таамаглал

$\rho_1 \geq \rho_2 \geq \ldots \geq \rho_k$ каноник корреляцын коэффициентуудын эхний s ширхэг нь л тэгээс ялгаатай харин үлдэх $k-s$ ширхэг нь тэгтэй тэнцүү гэсэн таамаглал үнэн бол

$$\chi^2 = -\{n - (p+q+3)/2\} \ln \prod_{i=s+1}^{k} (1-\rho_i^2)$$

статистик n буюу түүврийн хэмжээ хүрэлцээтэй их үед $(p-s+1)(q-s+1)$ чөлөөний зэрэг бүхий хи-квадрат тархалттай байдаг. Өөрөөр хэлбэл

$$H_0 : \rho_s = 0, \rho_{s+1} = 0, \ldots, \rho_k = 0 \quad s \geq 1$$

тэг таамаглал үнэн бол

$$\chi^2 \sim \chi^2_{(p-s+1)(q-s+1)}$$

байна.

Жишээ 38. Өмнө авсан жишээний хувьд $H_0 : \rho_1 = 0, \rho_2 = 0$ болон $H_0 : \rho_1 \neq 0, \rho_2 = 0$ таамаглалуудыг $\alpha = 0.05$ итгэх түвшинд шалга.

Түүврийн хэмжээ $n = 12$, санамсаргүй векторуудын хэмжээс харгалзан $p = 2$, $q = 3$, каноник корреляцын коэффициентууд $\rho_1 = \lambda_1^{1/2} = 0.84$ ба $\rho_2 = \lambda_2^{1/2} = 0.62$ байгааг анхаарвал уг таамаглалуудыг дараах байдлаар шалгана.

1. $H_0 : \rho_1 = 0, \rho_2 = 0$

   $$-\{12 - (2+3+3)/2\} \ln\{(1-0.84^2)(1-0.62^2)\} = 13.95$$

   p-утга $= 1 - \chi^2_6(13.95) = 0.03 < \alpha = 0.05$ тул тэг таамаглал няцаагдана.

2. $H_0 : \rho_1 \neq 0, \rho_2 = 0$ таамаглалын хувьд $s = 1$ байна.

   $$-\{12 - (2+3+3)/2\} \ln(1-0.62^2) = 3.92$$

   p-утга $= 1 - \chi^2_2(3.92) = 0.14 > \alpha = 0.05$ тул тэг таамаглалыг няцаах үндэслэлгүй.

Лекц XVI
Хамтын тархалтын холбоо хамаарлын шинжилгээ

1 Хамтын тархалтын холбоо хамаарлын шинжилгээ

Хамтын тархалтын холбоо хамаарлын шинжилгээ[14]

[14] Correspondence Analysis

Үүгээр чанарын хувьсагчдаас тогтох санамсаргүй векторын хамтын тархалтын хүснэгтэд тулгуурлан эдгээр хувьсагчдын холбоо хамаарлыг ангиудынх нь түвшинд задлан шинжилдэг.

Нүд	Үс			
	хар	хүрэн	улаан	цайвар
хүрэн	32	53	10	3
цэнхэр	11	50	10	30
бор	10	25	7	5
ногоон	3	15	7	8

Хүснэгт 11: Эрэгтэй хүний нүд болон үсний өнгө хувьсагчдын хамтын тархалт

Жишээ 39. Нүд болон үсний өнгөний уялдаа холбоог задлан шинжил.

Хамтын тархалтын хүснэгтийг R програмд матриц байдлаар оруулах

```
N <- matrix(
  data = c(
    32, 53, 10,  3,
    11, 50, 10, 30,
    10, 25,  7,  5,
     3, 15,  7,  8
  ), nrow = 4, ncol = 4, byrow = TRUE,
  dimnames = list(
    "Eye Color" = c("Brown", "Blue", "Hazel", "Green"),
    "Hair Color" = c("Black", "Brown", "Red", "Blond")
  )
)
```

Шинжилгээ хийх

```
CA <- ca::ca(N)
```

2 Хамааралгүй байх тухай таамаглал шалгах Хи-квадрат шинжүүр

Хамааралгүй байх тухай таамаглал шалгах χ^2 шинжүүр

X_1	X_2			Нийлбэр
	b_1	$\dots$	b_n	
a_1	n_{11}	$\dots$	n_{1n}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_m	n_{m1}	$\dots$	n_{mn}	$n_{m\cdot}$
Нийлбэр	$n_{\cdot 1}$	$\dots$	$n_{\cdot n}$	$n_{\cdot\cdot}$

Хүснэгт 12: Хоёр хэмжээст ангилсан өгөгдлийн хамтын давтамжийн хүснэгт

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(n_{ij}-\nu_{ij})^2}{\nu_{ij}} \sim \chi^2_{(m-1)(n-1)}$$

Энд $\nu_{ij} = \dfrac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$ байна.

Жишээ болгон авсан өгөгдлийг ашиглаж хи-квадрат шинжүүрээр эрэгтэй хүний нүдний өнгө ба үсний өнгө хамааралгүй гэсэн тэг таамаглалыг дараах байдлаар шалгаж болно.

```
chisq.test(x = N, correct = FALSE)
```

Эндээс дараах үр дүн гарна.

```
    Pearson's Chi-squared test

data:  N
X-squared = 41.28, df = 9, p-value = 4.447e-06
```

Ийнхүү p-утга $= 4.447 \cdot 10^{-6} < \alpha = 0.05$ байх тул эрэгтэй хүний нүдний өнгө ба үсний өнгө хамааралгүй гэсэн тэг таамаглалыг няцаана. Тэгвэл эдгээр хувьсагчдын чухам аль анги нөгөө хувьсагчид илүү нөлөөтэй бас ямар ангиуд илүү уялдаа холбоотойг хэрхэн тогтоох вэ?

3 Нөхцөлт тархалт болон тухайн тархалт хоорондын хи-квадрат зай буюу ангийн нөлөө

Нөхцөлт тархалт болон тухайн тархалт хоорондын хи-квадрат зай буюу ангийн нөлөө

Жишээний адилаар хувьсагчид хамааралтай байх үед тэдгээр хувьсагчдын холбоо хамаарлыг тайлбарлах шаардлага тулгарна. Тухайлбал нэг хувьсагч дээрх нөгөө хувьсагчийн аль нэг ангийн нөлөөг хэмжихдээ эхний хувьсагчийн тухайн тархалт болон сонгож авсан ангийн нөхцөл дэх нөхцөлт тархалт хоорондын хи-квадрат зайг ашигладаг. Ийм байдлаар тухайн нэг хувьсагчид нөгөө хувьсагчийн чухам аль анги илүү нөлөөтэй вэ гэсэн асуултад хариулдаг. Энэхүү зайг бодох томьёо болон уг зай χ^2 хи-квадрат статистиктай хэрхэн уялддаг талаар хичээлийн сурах бичгээс үзнэ үү. $\chi^2/n_{\cdot\cdot}$ статистик нь анги нэг бүрчлэн олсон ийм зайнуудын жинлэсэн нийлбэр юм. Иймд хэрэв $\chi^2/n_{\cdot\cdot}$ статистикийн утга бага бол түүний бүрэлдэхүүн хэсэг болох тэдгээр зай бүр бага утгатай болно. Харин $\chi^2/n_{\cdot\cdot}$ статистик их утгатай бол уг зайнуудын ядаж нэг нь их утгатай болно. Энэ нь хувьсагчдын хамааралгүй байдлыг хамтын давтамжийн хүснэгтийн чухам хаана үгүйсгэж байгааг тодорхойлох боломж олгоно.

Тус зайнууд нь `ca()` функцийн буцаах утгын `rowdist` болон `coldist` элементүүдэд хадгалагдана.

- Мөр буюу нүдний өнгө хувьсагчийн (Brown, Blue, Hazel, Green) анги тус бүрийн нөлөө

  ```
  CA$rowdist
  ```

```
0.4392956 0.3909005 0.1679889 0.4122632
```

Hazel буюу бор нүд холбоо хамааралд бага нөлөө үзүүлж байна.

- Багана буюу үсний өнгө хувьсагчийн (Black, Brown, Red, Blond) анги тус бүрийн нөлөө

```
CA$coldist
```

```
0.49975146 0.05697079 0.30858053 0.71615712
```

Blond буюу цайвар үс хамааралд хүчтэй нөлөөлж байна.

4 Хамтын тархалт болон хамааралгүйн нөхцөл дэх хамтын тархалт хоорондын хи-квадрат зай

Хамтын тархалт болон хамааралгүйн нөхцөл дэх хамтын тархалт хоорондын хи-квадрат зай

$P(X_1 = a_i, X_2 = b_j) = p_{ij}$ буюу X_1 болон X_2 санамсаргүй хувьсагчдын хамтын тархалт ба уг хувьсагчид хамааралгүй гэсэн нөхцөл дэх $P(X_1 = a_i)P(X_2 = b_j) = p_{i\cdot}p_{\cdot j}$ хамтын тархалт хоорондын

$$\chi^2_{X_1X_2} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$$

хи-квадрат зай нь тус хоёр хувьсагч хамааралгүйн нөхцөлд хэр ойр эсвэл хол байгааг илтгэнэ. Нөгөө талаас уг зайг i дүгээр мөр ба j дүгээр баганын огтлолцолд байх элемент нь $c_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}$ байдаг $C = (c_{ij})_{i=1,\ldots,m,\ j=1\ldots,n}$ матрицаар

$$\chi^2_{X_1X_2} = \mathrm{tr}(CC^T)$$

гэж илэрхийлж болно. c^2_{ij} буюу хи-квадрат зайн бүрэлдэхүүн хэсгүүдийг *инерц* гэдэг.

Хамтын тархалт болон хамааралгүйн нөхцөл дэх хамтын тархалт хоорондын хи-квадрат зайн хувийн утгын задаргаа

CC^T матриц X_1 болон X_2 чанарын хувьсагчдын хувьд ковариацийн матрицын үүрэг гүйцэтгэнэ. Иймд уг матрицыг гол хэсгийн шинжилгээнийх шиг задалж хувьсагчдын холбоо хамаарлыг гүнзгийрүүлэн шинжлэх боломжтой. CC^T матрицын хувийн утгуудыг $\lambda_1, \ldots, \lambda_r$ гэе. Энд $r = \mathrm{rank}(C)$. Тэгвэл

$$\chi^2_{X_1X_2} = \mathrm{tr}(CC^T) = \sum_{k=1}^{r}\lambda_k$$

задаргаа хүчинтэй байна. Иймхүү X_1 болон X_2 хувьсагчдын хамаарлын хэмжээс болох χ^2 зайг r ширхэг ортогонал буюу хамааралгүй факторт задаллаа. Факторуудын инерцийг CC^T матрицын $\lambda_1, \ldots, \lambda_r$ хувийн утгууд илэрхийлнэ.

CC^T матрицын $\lambda_1, \ldots, \lambda_r$ хувийн утгуудын язгуур буюу C матрицын $\lambda_1^{1/2}, \ldots, \lambda_r^{1/2}$ сингуляр утгууд нь `ca()` функцийн буцаах утгын `sv` элементэд агуулагдана.

```
CA$sv^2
```

```
0.1342877554 0.0132751891 0.0003950798
```

Иймд фактор бүрийн инерц нийт инерцийн 90.76, 8.97, 0.27 хувийг тус тус эзэлнэ.

Мөн (c_{ij}^2) матрицын мөр, баганын дагуу нийлбэр авч мөр болон баганын инерцийг олж болно. Эдгээр нь `ca()` функцийн буцаах утгын `rowinertia` болон `colinertia` элементүүдэд тус тус агуулагдана.

Хамтын тархалт болон хамааралгүйн нөхцөл дэх хамтын тархалт хоорондын хи-квадрат зай ба хи-квадрат шинжүүрийн статистик

Хамтын тархалтыг хамтын давтамжаар $\hat{p}_{ij} = \frac{n_{ij}}{n_{..}}$ гэж үнэлээд нийт инерцийн илэрхийллийг ахин бичвэл

$$\chi^2_{X_1X_2} = \sum_{k=1}^{r} \lambda_k = \mathrm{tr}(CC^T) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{1}{n_{i.}n_{.j}} \left(n_{ij} - \frac{n_{i.}n_{.j}}{n_{..}} \right)^2 = \frac{\chi^2}{n_{..}}$$

болж хи-квадрат шинжүүрийн статистиктай уялдана.

5 Факторууд дээрх мөр, баганын проекц

Факторууд дээрх мөр, баганын проекц

Чанарын хувьсагчийн ангиуд буюу хамтын давтамжийн хүснэгтийн мөр баганын уялдаа холбоог тогтоохын тулд өмнөх хэсэгт дурдсан CC^T матрицын хувьд хийсэн гол хэсгийн шинжилгээгээр олсон факторууд дээрх мөр болон баганын жинлэсэн проекцыг авч үздэг. CC^T матрицын хувийн утгын задаргаатай уялдах C матрицын сингуляр утгын задаргаа

$$C = \Gamma\Lambda\Delta^T$$

хэлбэртэй байна. Энд буй $\Lambda = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ нь CC^T болон C^TC матрицын тэг биш $r = \mathrm{rank}(C)$ ширхэг $\lambda_1, \ldots, \lambda_r$ хувийн утгуудын язгуураас тогтох диагональ матриц, Γ болон Δ бол харгалзан CC^T болон C^TC матрицын эхний r ширхэг хувийн утгад харгалзах хувийн векторуудаас тогтох матриц юм.

P_{X_1} болон P_{X_2} тухайн тархалтуудын Δ болон Γ хувийн векторууд дээрх C матрицаар жинлэсэн проекц

$$S^T = P_{X_1}^{-1/2} C\Delta$$
$$V^T = P_{X_2}^{-1/2} C\Gamma$$

нь X_1 болон X_2 хувьсагчдын анги тус бүр факторууд дээр хаана байрлахыг илэрхийлэх бөгөөд эдгээрийг *гол координат* гэдэг. Энд S^T болон V^T нь хөрвөсөн матрицууд юм.

Дээрх гол проекцоос гадна $S_{\mathrm{std}}P_{X_1}S_{\mathrm{std}}^T = I_r$ ба $V_{\mathrm{std}}P_{X_2}V_{\mathrm{std}}^T = I_r$ чанартай

$$S_{\mathrm{std}} = \Lambda^{-1}P_{X_1}^{-1/2}C\Delta$$
$$V_{\mathrm{std}} = \Lambda^{-1}P_{X_2}^{-1/2}C\Gamma$$

стандарт проекц авч үздэг. Энд I_r бол r хэмжээст нэгж матриц юм. Энэ тохиолдолд S_{std} хувьсагч X_1 хувьсагчтай мөн V_{std} хувьсагч X_2 хувьсагчтай хамааралгүй юм. S_{std}^T болон V_{std}^T матрицуудаар илэрхийлэгдэх координатуудыг *стандарт координат* гэдэг.

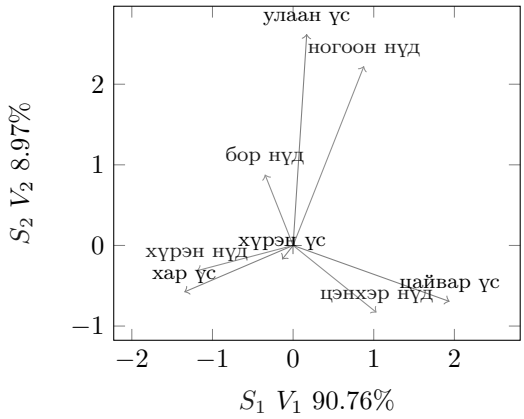
Мөр буюу X_1 хувьсагчийн ангиуд бас багана буюу X_2 хувьсагчийн ангиудын стандарт координатууд `ca()` функцийн буцаах утгын `rowcoord` болон `colcoord` элементүүдэд агуулагдана.

```
CA$rowcoord
```

```
           Dim1       Dim2        Dim3
Brown -1.194305 -0.3141723  0.56733681
Blue   1.034212 -0.8310791 -0.04571967
Hazel -0.348969  0.8795746 -2.01015903
Green  0.878425  2.2238748  1.31806526
```

```
CA$colcoord
```

```
            Dim1       Dim2       Dim3
Black -1.3496098 -0.5777888  1.3516125
Brown -0.1361648 -0.1734383 -0.9499617
Red    0.1689698  2.6220451  0.5497371
Blond  1.9414076 -0.6954714  0.9013731
```



Зураг 26: Эхний хоёр фактор дээрх мөр болон баганын стандарт проекц

Ийм диаграмм байгуулахын тулд `plot(CA)` байдалтай тушаал өгнө.

Цэгэн диаграмм дээрх зарим цэг бусадтай нь харьцуулахад ойролцоо байрлах нь тэдгээрт харгалзах ангиудыг бусдаасаа илүү холбоо хамааралтай болохыг илтгэнэ. Тодруулбал

- X_1 (мөр) хувьсагчийн ангиудад харгалзах цэгүүд ойролцоо байх нь тэдгээрийн X_2 хувьсагчийн нөхцөл дэх тархалт нь төстэй болохыг
- X_2 (багана) хувьсагчийн ангиудад харгалзах цэгүүд ойролцоо байх нь тэдгээрийн X_1 хувьсагчийн нөхцөл дэх тархалт нь төстэй болохыг

- X_1 (мөр) хувьсагчид харгалзах нэг цэг X_2 (багана) хувьсагчид хувьсагчид харгалзах өөр нэг цэгтэй ойролцоо байх нь эдгээр хувьсагчийн тухайн хоёр анги өөр хоорондоо холбоо хамааралтай болохыг

тус тус илтгэн харуулна.

Хэрэв санамсаргүй хувьсагч дараалсан хэмжээстэй бол ангиудынх нь эрэмбэ дарааллын дагуу тэдгээрт харгалзах цэгүүдийг шугамаар холбож зурсанаар зарим нэмэлт зүй тогтол ажиглах боломжтой. Үүний тулд `plot()` функцийн `lines` аргументаар `TRUE` утга дамжуулна.

Мөр ба багана дахь анги тус бүрийн фактор дээрх оролцоо

Чанар 17. s_k болон v_k нь k дугаар факторт харгалзах гол координатууд буюу харгалзах матрицын мөрийн элементүүдийг төлөөлсөн санамсаргүй хувьсагч байг. Тэгвэл дараах адилтгал биелнэ.

$$Es_k = 0 \qquad \operatorname{cov}(s_k) = \sum_{i=1}^{m} p_{i\cdot} s_{ki}^2 = \lambda_k$$

$$Ev_k = 0 \qquad \operatorname{cov}(v_k) = \sum_{j=1}^{n} p_{\cdot j} v_{kj}^2 = \lambda_k$$

Энд s_{ki} болон v_{kj} харгалзан s_k болон v_k санамсаргүй хувьсагчийн i болон j дүгээр утга юм.

Дээрх чанараас дараах мөрдлөгөө гарна.

Мөрдлөгөө 3. λ_k буюу

$$\frac{\lambda_k}{\sum_{l=1}^{r} \lambda_l}$$

харьцаа нь хи-квадрат зайн задаргаа дахь k дугаар факторын ковариацийн хувь хэмжээг илэрхийлнэ.

Өмнөх чанарт үндэслэн зохиосон

$$C_{X_1}(i, s_k) = \frac{p_{i\cdot} s_{ki}^2}{\lambda_k}, \quad i = 1, \ldots, m,\ k = 1, \ldots, r$$

$$C_{X_2}(j, v_k) = \frac{p_{\cdot j} v_{kj}^2}{\lambda_k}, \quad j = 1, \ldots, n,\ k = 1, \ldots, r$$

харьцаанууд харгалзан s_k фактор дахь i дүгээр мөрийн *оролцоо* (ковариацад эзлэх хувь) болон v_k фактор дахь j дүгээр баганын оролцоог илэрхийлнэ.

Жишээний хувьд нийт инерцийн 90.76 хувийг дангаараа эзлэх $k = 1$ дүгээр фактор дахь X_1 хувьсагчийн ангиуд буюу мөрийн оролцоо

$$C_{X_1}(\,\cdot\,, s_1) = (0.501, 0.387, 0.020, 0.091)$$

харин X_2 хувьсагчийн ангиуд буюу баганын оролцоо

$$C_{X_2}(\,\cdot\,, v_1) = (0.365, 0.009, 0.003, 0.621)$$

байна. Иймд хувьсагчдын хамааралд 1 дүгээр мөр буюу "хүрэн нүд" болон 4 дүгээр багана буюу "цайвар үс" гэсэн ангиуд голлох нөлөөтэй гэж дүгнэнэ.

Уг оролцоог `summary(CA)` байдлаар олж болно.

Ном зүй

[1] Г.Махгал, Ш.Мөнгөнсүх *Олон хэмжээст өгөгдлийн статистик шинжилгээ 2017, ISBN 978-99978-1-648-1.*