Олон хэмжээст өгөгдлийн статистик шинжилгээ

лекц

Г.Махгал

₮ 2021/1/27

Хичээлийн веб хуудас

хаяг www.magadlal.com/courses/5.html

агуулга видео лекц, семинар болон гэрийн даалгаврын бодлого, R хэл дээрх дасгал ажил, бие даалтын ажлын удирдамж, дүгнэх журам

Хичээлийн сурах бичиг

нэр Олон хэмжээст өгөгдлийн статистик шинжилгээ

зохиогч Г. Махгал, Ш. Мөнгөнсүх

хэвлэсэн он 2017

Ашиглах програм

үндсэн програм R

туслах програм RStudio

ном сурах бичиг www.magadlal.com/books/id-2.html
www.magadlal.com/books/id-4.html

Агуулга

1. Удиртгал
2. Олон хэмжээст тархалт
3. Олон хэмжээст хэвийн тархалт I
4. Олон хэмжээст хэвийн тархалт II
5. Олон хэмжээст хэвийн тархалт III
6. Олон хэмжээст хэвийн тархалт IV
7. Гол хэсгийн шинжилгээ
8. Факторын шинжилгээ - I хэсэг
9. Факторын шинжилгээ - II хэсэг
10. Кластерын шинжилгээ - I хэсэг
11. Кластерын шинжилгээ - II хэсэг
12. Дискриминантын шинжилгээ
13. Олон хэмжээст координатын шинжилгээ
14. Конжойнт шинжилгээ
15. Каноник корреляцын шинжилгээ

Лекц I

Удиртгал

Удиртгал сэдвийн агуулга

1 Хичээлийн агуулга

2 Үндсэн шинжилгээнүүд

3 Шаардагдах суурь мэдлэг

Хэсэг 1

Хичээлийн агуулга

Хичээлийн агуулга

Үндсэн шинжилгээнүүд
1. Гол хэсгийн шинжилгээ
2. Факторын шинжилгээ
3. Кластерын шинжилгээ
4. Дискриминантын шинжилгээ
5. Олон хэмжээст координатын шинжилгээ
6. Конжойнт шинжилгээ
7. Каноник корреляцын шинжилгээ
8. Хамтын тархалтын холбоо хамаарлын шинжилгээ

Бусад
9. Олон хэмжээст хэвийн тархалт
10. Копула

Хэсэг 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[1]

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

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

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

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

[1] Principal Component Analysis

Факторын шинжилгээ[2]

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

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

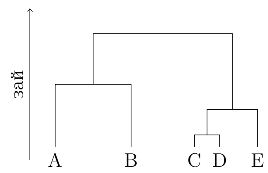
$$L = ?, \quad F = ?$$

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

[2] Factor Analysis

Кластерын шинжилгээ[3]

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

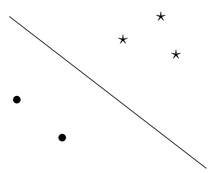


Зураг: Дендрограмм

[3] Cluster Analysis

Дискриминантын шинжилгээ[4]

Бүлгүүдийн заагийг олж тогтоох зэрэгт ашигладаг.

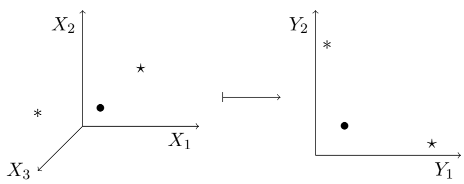


Зураг: Дискриминантын шулуун

[4] Discriminant Analysis

Олон хэмжээст координатын шинжилгээ[5]

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



Зураг: Хэмжээс багасгах

[5] Multidimensional Scaling

Конжойнт шинжилгээ[6]

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

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

(a) Өгөгдөл

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

(b) Үр дүн

Хүснэгт: Шинжилгээний өгөгдөл ба эцсийн үр дүн

[6] Conjoint Analysis

Каноник корреляцын шинжилгээ[7]

Хоёр санамсаргүй вектор хоорондын холбоо хамаарлыг хэмжихэд ашиглана.

$$\rho(X, Y)$$

энд

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

[7] Canonical Correlation Analysis

Хамтын тархалтын холбоо хамаарлын шинжилгээ[8]

Чанарын хувьсагчдын уялдаа холбоог задлан шинжлэхэд ашигладаг.

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

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

[8] Correspondence Analysis

Хэсэг 3

Шаардагдах суурь мэдлэг

Шаардагдах мэдлэг, чадвар

1. Матриц, түүн дээрх үйлдэл, чанар, хувийн утга болон хувийн вектор
2. Олон хэмжээст тархалт, тухайн тархалт, нөхцөлт тархалт
3. Момент, ковариацын 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.

Тодорхойлолт (Хувийн утга болон хувийн вектор)

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

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$values    # eigen values
eig$vectors   # eigen vectors
```


Олон хэмжээст тархалт

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

Тодорхойлолт (Хамтын тархалтын функц)

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

Тодорхойлолт (Хамтын нягтын функц)

$f_X(x) \geq 0$:

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

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

Тухайн тархалт

$X = (X_1, X_2)^T$ энд $X_1 \in \mathbb{R}^k$ ба $X_2 \in \mathbb{R}^{p-k}$.

Тодорхойлолт (Тухайн тархалтын функц)

$$F_{X_1}(x_1) = P(X_1 < x_1) = F_X(x_1, \ldots, x_k, \infty, \ldots, \infty)$$

Тодорхойлолт (Тухайн нягтын функц)

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

Нөхцөлт тархалт

$X_1 = x_1$ гэе.

Тодорхойлолт (Нөхцөлт нягтын функц)

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

Тодорхойлолт (Хамааралгүй)

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

бол X_2 хувьсагчийг X_1 хувьсагчаас хамааралгүй гэнэ.

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. $a, b \in \mathbb{R}$ 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 ковариацын матриц

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

Тодорхойлолт (Ковариацын матриц)

$$DX = \Sigma = \Sigma_{XX} = E(X-\mu)(X-\mu)^T$$
$$= \begin{pmatrix} \operatorname{cov}(X_1,X_1) & \operatorname{cov}(X_1,X_2) & \dots & \operatorname{cov}(X_1,X_p) \\ \operatorname{cov}(X_2,X_1) & \operatorname{cov}(X_2,X_2) & \dots & \operatorname{cov}(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p,X_1) & \operatorname{cov}(X_p,X_2) & \dots & \operatorname{cov}(X_p,X_p) \end{pmatrix}$$

Чанар

$$\Sigma^T = \Sigma$$

Хоёр өөр санамсаргүй векторын ковариацын матриц

Хоёр өөр санамсаргүй векторын ковариацын матриц дараах хэлбэртэй байна.

$$\operatorname{cov}(X,Y) = \Sigma_{XY} = E(X-\mu)(Y-\nu)^T$$
$$= \begin{pmatrix} \operatorname{cov}(X_1,Y_1) & \operatorname{cov}(X_1,Y_2) & \dots & \operatorname{cov}(X_1,Y_q) \\ \operatorname{cov}(X_2,Y_1) & \operatorname{cov}(X_2,Y_2) & \dots & \operatorname{cov}(X_2,Y_q) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p,Y_1) & \operatorname{cov}(X_p,Y_2) & \dots & \operatorname{cov}(X_p,Y_q) \end{pmatrix}$$

Чанар

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 ковариацын матриц олох

R програм дээрх матриц, датафрейм, тиббл зэрэг хэлбэртэй олон хэмжээст өгөгдлийн хувьд түүврийн дундаж утгын вектор болон түүврийн ковариацын 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

Олон хэмжээст тархалт сэдвийн агуулга

1 Санамсаргүй векторын математик дундаж олох
2 Ковариацын матриц
3 Нөхцөлт математик дундаж
4 Нөхцөлт ковариац буюу тухайн ковариац
5 Санамсаргүй векторын хувиргалт

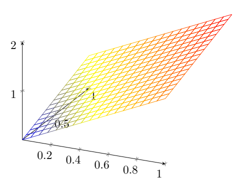
Хэсэг 1

Санамсаргүй векторын математик дундаж олох

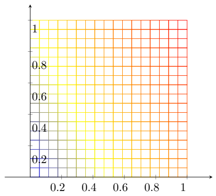
Жишээ болгон авч үзэх хоёр хэмжээст тархалт

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

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



(a) нягтын функцийн график



(b) эгц дээрээс нь

Зураг: $(*)$ нягтын хэлбэр

$(*)$ 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)
    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

$$f_{X_1}(x_1)=\int_{-\infty}^{+\infty} f_{X_1,X_2}(x_1,x_2)dx_2=\int_0^1\frac{4x_1+2x_2}{3}dx_2=\frac{4x_1+1}{3}$$

$$EX_1=\int_{-\infty}^{+\infty} x_1 f_{X_1}(x_1)dx_1=\int_0^1 x_1\frac{4x_1+1}{3}dx_1=\frac{11}{18}\approx 0.611$$

$$f_{X_2}(x_2)=\int_{-\infty}^{+\infty} f_{X_1,X_2}(x_1,x_2)dx_1=\int_0^1\frac{4x_1+2x_2}{3}dx_1=\frac{2+2x_2}{3}$$

$$EX_2=\int_{-\infty}^{+\infty} x_2 f_{X_1}(x_1)dx_2=\int_0^1 x_2\frac{2+2x_2}{3}dx_2=\frac{5}{9}\approx 0.555$$

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

Хэсэг 2

Ковариацын матриц

Ковариацын матрицын чанар

Чанар

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

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

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

Баталгаа

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

□

[1] $\operatorname{cov}(X,Y)=E(X-EX)(Y-EY)^T$

Санамсаргүй векторын ковариацын матриц олох

$$DX_1=EX_1^2-(EX_1)^2=\int_{-\infty}^{+\infty}x_1^2f_{X_1}(x_1)dx_1-(EX_1)^2$$

$$=\int_0^1 x_1^2\frac{4x_1+1}{3}dx_1-\frac{11^2}{18^2}=\frac{4}{9}-\frac{121}{324}=\frac{23}{324}\approx 0.071$$

$$DX_2=\frac{13}{162}\approx 0.080$$

$$\operatorname{cov}(X_1,X_2)=E(X_1X_2)-EX_1EX_2=\int_{\mathbb{R}^2}x_1x_2f(x_1,x_2)dx_1dx_2$$

$$-EX_1EX_2=\int_0^1\int_0^1 x_1x_2\frac{4x_1+2x_2}{3}dx_1dx_2-\frac{11}{18}\cdot\frac{5}{9}$$

$$=\frac{1}{3}-\frac{55}{162}=-\frac{1}{162}\approx -0.006$$

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

Санамсаргүй векторын корреляцын матриц олох

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

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

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.

Хэсэг 3

Нөхцөлт математик дундаж

Нөхцөлт математик дундаж

Тодорхойлолт

$$E(X_2|X_1 = x_1) = \int x_2 f_{X_2|X_1}(x_2|x_1) dx_2$$

Чанар

1. $E(E(X_2|X_1)) = EX_2$ (Бүтэн дунджийн томьёо)
2. $E(X_2 + X_3|X_1) = E(X_2|X_1) + E(X_3|X_1)$
3. X_1 ба X_2 хамааралгүй бол $E(X_2|X_1) = EX_2$
4. $E(\varphi(X_1)X_2|X_1) = \varphi(X_1)E(X_2|X_1)$

Жишээ

Бүтэн дунджийн томьёог батал.

Баталгаа

$$\begin{aligned} E(E(X_2|X_1)) &= \int E(X_2|X_1) f(x_1) dx_1 \\ &= \int \left(\int x_2 \cdot f(x_2|x_1) dx_2 \right) f(x_1) dx_1 \\ &= \int \left(\int x_2 \cdot f_{x_1,x_2}(x_1, x_2) dx_1 \right) dx_2 \\ &= \int x_2 \left(\int f(x_1, x_2) dx_1 \right) dx_2 \\ &= \int x_2 \cdot f(x_2) dx_2 \\ &= EX_2 \end{aligned}$$

□

Нөхцөлт математик дундаж олох

Жишээ

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.

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

Нөхцөлт 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 хувьд хамаарлыг $E(X_2|X_1 = x_1) = a + bx_1$ байдлаар авч үздэг. Тэгвэл шугаман болон шугаман бус загварын хооронд хэр ялгаа байх бол?
Шугаман загварыг ойролцоогоор үнэлэхдээ симуляцийн аргаар өмнө гаргаж авсан хиймэл өгөгдөл болон `lm()` функц ашиглая.

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

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")
```


Хэсэг 4

Нөхцөлт ковариац буюу тухайн ковариац

Нөхцөлт ковариац буюу тухайн ковариац

Тодорхойлолт

$$\begin{aligned} &\operatorname{cov}(X_2|X_1 = x_1) = \\ &\quad = E((X_2 - E(X_2|X_1 = x_1))(X_2 - E(X_2|X_1 = x_1))^T|X_1 = x_1) \end{aligned}$$

Чанар

1. $\operatorname{cov}(X_2) = E(\operatorname{cov}(X_2|X_1)) + \operatorname{cov}(E(X_2|X_1))$ (Бүтэн ковариацын томьёо)
2. $\operatorname{cov}(X_2|X_1) = E(X_2 X_2^T|X_1) - E(X_2|X_1)E(X_2^T|X_1)$

Жишээ

Бүтэн ковариацын томьёог батал.

Баталгаа

$$
\begin{aligned}
&E(\mathrm{cov}(X_2|X_1)) + \mathrm{cov}(E(X_2|X_1)) = \\
&\quad = E\left\{E(X_2X_2^T|X_1) - E(X_2|X_1)E(X_2^T|X_1)^{2}\right\} + \\
&\quad\quad + \left[E\left\{E(X_2|X_1)E(X_2^T|X_1)\right\} - EX_2EX_2^{T\,3}\right]^{4} \\
&\quad = E(X_2X_2^T) - EX_2EX_2^T \\
&\quad = \mathrm{cov}(X_2)
\end{aligned}
$$

□

[2] нөхцөлт ковариацын 2 дугаар чанар
[3] бүтэн дунджийн томьёо
[4] ковариацын математик дунджаарх задаргаа

Тухайн корреляц

Тодорхойлолт (Тухайн корреляц буюу нөхцөлт корреляц)

$$
\rho(X_2, X_3|X_1 = x_1) = \\
= \frac{\mathrm{cov}(X_2, X_3|X_1 = x_1)}{\sqrt{\mathrm{cov}(X_2|X_1 = x_1)\cdot\mathrm{cov}(X_3|X_1 = x_1)}}
$$

Хэсэг 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

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

Энд $J = \left(\frac{\partial x_i(y)}{\partial y_j}\right)_{i,j=1,\dots,p}$ 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.

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

Жишээ

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 $f_Y(y) = \mathrm{abs}(|A|^{-1})f_X(A^{-1}(y - b))$ томьёо ашиглана. Энд $|A| = -2$, $\mathrm{abs}(|A|^{-1}) = \frac{1}{2}$, $A^{-1} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$ тул

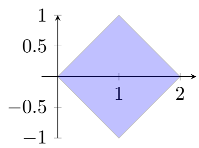
$$
\begin{aligned}
f_Y(y) &= \frac{1}{2} f_X\left\{\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right\} = \frac{1}{2} f_X\left\{\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2}\right\} \\
&= \frac{1}{2}\cdot\frac{4\cdot\frac{y_1+y_2}{2} + 2\cdot\frac{y_1-y_2}{2}}{3}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

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: Y санамсаргүй векторын авах утгууд

```r
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

Олон хэмжээст хэвийн тархалт I сэдвийн агуулга

1 Тодорхойлолт
2 Хувиргалт ба загварчлал
3 Геометр агуулга
4 Копула

Хэсэг 1

Тодорхойлолт

Тодорхойлолт

Сэргээн санах нь (Нэг хэмжээст хэвийн тархалт)

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

Тодорхойлолт

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

нягттай тархалтыг *олон хэмжээст хэвийн тархалт* гэнэ.

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

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

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

$$\Sigma = \operatorname{cov}(X) = E(X-\mu)(X-\mu)^T$$

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

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 ковариацын матриц

Тархалтын функц

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

- lower доод хязгаар
- upper дээд хязгаар
- mean дундаж утгын вектор
- sigma ковариацын матриц

Жишээ

$\mu = (5, -4)$ дундаж утгын вектор болон $\Sigma = \begin{pmatrix}4 & -1\\ -1 & 1\end{pmatrix}$ ковариацын 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

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

болно.

Хоёр хэмжээст хэвийн тархалт

ρ корреляцтай X_1 ба X_2 санамсаргүй хувьсагчдаас тогтох $X = (X_1, X_2)$ 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

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

болно.

Жишээ

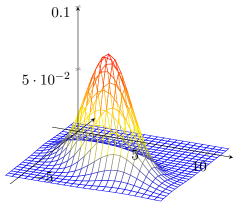
$\mu = (5, -4)$ дундаж утгын вектор болон $\Sigma = \begin{pmatrix}4 & -1\\ -1 & 1\end{pmatrix}$ ковариацын матриц бүхий хоёр хэмжээст хэвийн тархалтын хамтын нягтын илэрхийллийг бичиж, графикийг нь зур.

$\mu = (\mu_1, \mu_2) = (5, -4)$ 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

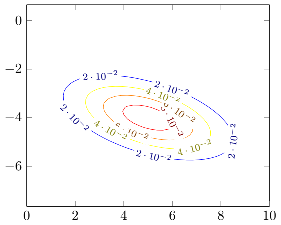
$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\sqrt{3}\pi} \exp\left\{-\frac{1}{6}\left[(x_1-5)^2 + 2(x_1-5)(x_2+4) + 4(x_2+4)^2\right]\right\}$$

болно. Харин графикийг нь дараагийн слайд дээр байгуулж үзүүлье.

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



нягтын функцийн график



нягтын функцийн түвшний шугам

Хэсэг 2

Хувиргалт ба загварчлал

Махаланобисын хувиргалт

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

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

байдлаар хувиргавал олон хэмжээст стандарт хэвийн тархалттай өөрөөр хэлбэл

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

болно. Энд $Y = (Y_1, \dots, Y_p)$, 0 нь тэг вектор, I_p нь p хэмжээст нэгж матриц юм. Иймд $Y_1, \dots, Y_p$ 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 загварчлал

$Y \sim N_p(0, I_p)$ буюу $Y = (Y_1, \dots, Y_p)$, $Y_1, \dots, Y_p$ хувьсагчид хамааралгүй бөгөөд нэг хэмжээст $N(0, 1)$ стандарт хэвийн тархалттай бол

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

болно.

$$E(\Sigma^{1/2}Y + \mu) = E(\Sigma^{1/2}Y) + E\mu = \Sigma^{1/2}EY + \mu = \mu$$

$$\begin{aligned}
&\operatorname{cov}(\Sigma^{1/2}Y + \mu)\\
&\quad = E(\Sigma^{1/2}Y + \mu - E(\Sigma^{1/2}Y + \mu))(\Sigma^{1/2}Y + \mu - E(\Sigma^{1/2}Y + \mu))^T\\
&\quad = E(\Sigma^{1/2}Y)(\Sigma^{1/2}Y)^T\\
&\quad = \Sigma^{1/2}E(YY^T)\Sigma^{1/2} \text{ [1]}\\
&\quad = \Sigma^{1/2} I_p \Sigma^{1/2}\\
&\quad = \Sigma
\end{aligned}$$

Олон хэмжээст хэвийн тархалттай санамсаргүй утга үүсгэх бэлэн функц

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

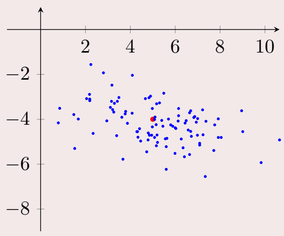
- **n** үүсгэх санамсаргүй утгын тоо
- **mean** дундаж утгын вектор
- **sigma** ковариацын матриц

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

- **n** үүсгэх санамсаргүй утгын тоо
- **mu** дундаж утгын вектор
- **Sigma** ковариацын матриц

Жишээ

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



Хувиргалт

Теорем

$X \sim N_p(\mu, \Sigma)$ ба A нь p эрэмбийн квадрат, үл бөхөх матриц; b нь p хэмжээст вектор бол $Y = AX + b$ санамсаргүй вектор p хэмжээст хэвийн тархалттай, өөрөөр хэлбэл,

$$Y \sim N_p(A\mu + b, A\Sigma A^T)$$

байна.

Баталгаа

- $f_Y(y) = \operatorname{abs}(|A|^{-1}) f_X(A^{-1}(y - b))$
- $f_X(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\}$
- $\left[(x - \mu)^T\Sigma^{-1}(x - \mu)\right]_{x = A^{-1}(y-b)} = (A^{-1}(y - b) - \mu)^T \Sigma^{-1}(A^{-1}(y - b) - \mu) = (A^{-1}(y - b) - A^{-1}A\mu)^T\Sigma^{-1}(A^{-1}(y - b) - A^{-1}A\mu) = (y - (A\mu + b))^T(A^{-1})^T$[2]$\Sigma^{-1}A^{-1}(y - (A\mu + b)) = (y - (A\mu + b))^T(A^T)^{-1}\Sigma^{-1}A^{-1}$[3]$(y - (A\mu + b)) = (y - (A\mu + b))^T(A\Sigma A^T)^{-1}(y - (A\mu + b))$

$\square$

[2] $(A^T)^{-1} = (A^{-1})^T$
[3] $(AB)^{-1} = B^{-1}A^{-1}$

Хэсэг 3

Геометр агуулга

Геометр агуулга

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

Чанар

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
   $$U = (X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_p \text{ [4]}$$
   байна.

[4] p чөлөөний зэрэгтэй хи-квадрат тархалт

Нэг дүгээр чанар илэрхий тул хоёр дугаар чанарыг л баталъя.

Баталгаа

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

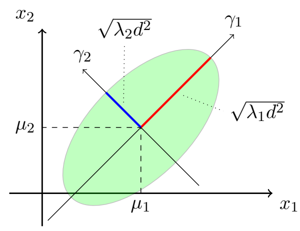
$$\begin{aligned} U &= (X-\mu)^T\Sigma^{-1}(X-\mu) \\ &= (X-\mu)^T\Sigma^{-1/2}\Sigma^{-1/2}(X-\mu) \\ &= (\Sigma^{-1/2}(X-\mu))^T\Sigma^{-1/2}(X-\mu)^{[5]} \\ &= Y^TY \\ &= \sum_{i=1}^{p} Y_i^{2\,[6]} \sim \chi^2_p \end{aligned}$$

□

[5] $Y = \Sigma^{-1/2}(X-\mu) \sim N_p(0,I)$

[6] $Y_i \sim N(0,1)$

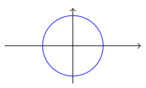
$(x-\mu)^T\Sigma^{-1}(x-\mu) = d^2$ эллипсоидийн их болон бага тэнхлэгүүдийн чиглэл болон хэмжээ



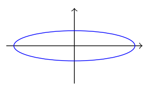
Энд λ_i болон γ_i нь Σ ковариацын матрицын хувийн утга болон хувийн вектор юм.

Мөн эллипсоидын "талбай" $S = \dfrac{2\pi^{p/2}}{p\Gamma(p/2)} d^p |\Sigma|^{1/2}$ байна.

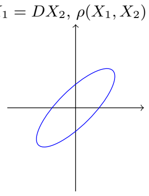
$(x-\mu)^T\Sigma^{-1}(x-\mu) = d^2$ эллипсоидийн хэлбэр болон их тэнхлэгийн чиглэл ба Σ матриц



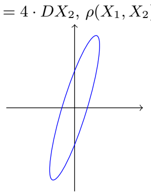
$DX_1 = DX_2,\ \rho(X_1,X_2) = 0$



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



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



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

Махаланобисын зай

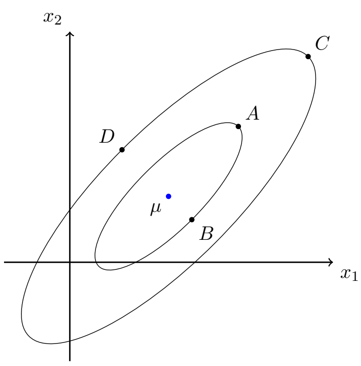
x цэгээс Σ ковариацын матрицтай тархалтын төв μ хүртэлх

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

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.



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

Хэсэг 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 ($p = 2$ 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. $u_1 \le v_1$ ба $u_2 \le v_2$ байх $\forall (u_1,u_2),(v_1,v_2) \in [0,1] \times [0,1]$: $C(v_1,v_2) - C(v_1,u_2) - C(u_1,v_2) + C(u_1,u_2) \ge 0$

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

Копула

Теорем (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

Олон хэмжээст хэвийн тархалт II сэдвийн агуулга

1 Шугаман хувиргалт
2 Нөхцөлт тархалт
3 Нөхцөлт дөхөлт
4 Параметрийн үнэлэлт

Хэсэг 1

Шугаман хувиргалт

Хувиргалт

Теорем

$X \sim N_p(\mu, \Sigma)$, $A_{(q\times p)}$, $c \in \mathbb{R}^q$, $q \le p$ 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 баталгаар нь бие даан үзнэ үү.

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

$$X \left\{ \begin{array}{c} \left.\begin{array}{c} X_1 \\ \vdots \\ X_r \end{array}\right\} X_1 \\ \left.\begin{array}{c} X_{r+1} \\ \vdots \\ X_p \end{array}\right\} X_2 \end{array} \right.$$

Өөрөөр хэлбэл $X_1 = (X_1, \ldots, X_r)$ ба $X_2 = (X_{r+1}, \ldots, X_p)$ гэе. Тэгвэл ковариацын матриц

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

блок матриц болно. Энд $\Sigma_{11} = \mathrm{cov}(X_1, X_1)$, $\Sigma_{22} = \mathrm{cov}(X_2, X_2)$, $\Sigma_{12} = \mathrm{cov}(X_1, X_2)$, $\Sigma_{21} = \Sigma_{12}^T = \mathrm{cov}(X_2, X_1)$ бөгөөд

$$X_1 \sim N_r(\mu_1, \Sigma_{11}) \quad X_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$$

байна.

Жишээ

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

$r = 2$ үед $X = (X_1, X_2, X_3)$ санамсаргүй вектор $X_1 = (X_1, X_2)$ ба $X_2 = (X_3)$ гэсэн хоёр дэд вектор болж задарна.
Ковариацын матриц нь дараах блок матриц болно.

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

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

Мөн $\mu_1 = (0, 0)$, $\mu_2 = (2)$ болно.

Теорем

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

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

бас

$$X_{2.1} = X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1$$

гэе. Тэгвэл X_1 ба $X_{2.1}$ хамааралгүй бөгөөд

$$X_1 \sim N_r(\mu_1, \Sigma_{11}), \qquad X_{2.1} \sim N_{p-r}(\mu_{2.1}, \Sigma_{22.1})$$

байна. Энд $\mu_{2.1} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1$, $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

Баталгаа

$$X_1 = \underbrace{\begin{pmatrix}I_r & 0\end{pmatrix}}_{A}\underbrace{\begin{pmatrix}X_1\\X_2\end{pmatrix}}_{X} = AX$$

$$X_{2.1} = X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1 = \underbrace{\begin{pmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I_{p-r}\end{pmatrix}}_{B}\underbrace{\begin{pmatrix}X_1\\X_2\end{pmatrix}}_{X} = BX$$

тул өмнөх теорем ёсоор $X_1 \sim N_r(\mu_1, \Sigma_{11})$ бас $X_{2.1} \sim N_{p-r}(\mu_{2.1}, \Sigma_{22.1})$ болно. Энд

$$\mu_{2.1} = B\mu + 0 = \begin{pmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I_{p-r}\end{pmatrix}\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix} = -\Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \mu_2$$
$$\Sigma_{22.1} = B\Sigma B^T = \begin{pmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I_{p-r}\end{pmatrix}\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}-\Sigma_{11}^{-1}\Sigma_{12}\\ I_{p-r}\end{pmatrix}$$
$$= \begin{pmatrix}0 & -\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22}\end{pmatrix}\begin{pmatrix}-\Sigma_{11}^{-1}\Sigma_{12}\\ I_{p-r}\end{pmatrix} = -\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22}$$

байна. ...

Баталгааны үргэлжлэл

X_1 болон $X_{2.1}$ санамсаргүй векторууд хамааралгүй болох нь

$$\operatorname{cov}(X_1, X_{2.1}) = \operatorname{cov}(AX, BX) = A\operatorname{cov}(X, X)B^T = A\Sigma B^T =$$
$$\left(\boxed{\begin{matrix}1&\dots&0\\\vdots&\ddots&\vdots\\0&\dots&1\end{matrix}}\ \boxed{\begin{matrix}0&\dots&0\\\vdots&\ddots&\vdots\\0&\dots&0\end{matrix}}\right)\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}(-\Sigma_{21}\Sigma_{11}^{-1})^T\\ \boxed{\begin{matrix}1&\dots&0\\\vdots&\ddots&\vdots\\0&\dots&1\end{matrix}}\end{pmatrix} =$$
$$\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\end{pmatrix}\begin{pmatrix}(-\Sigma_{21}\Sigma_{11}^{-1})^T\\ I_{p-r}\end{pmatrix} = \left(\Sigma_{11}(-\Sigma_{21}\Sigma_{11}^{-1})^T + \Sigma_{12}\right) =$$
$$\left(-\Sigma_{11}\Sigma_{11}^{-1}\Sigma_{12}{}^{[1]} + \Sigma_{12}\right) = 0_{r\times p-r}$$

□

[1] $\Sigma_{21} = (\Sigma_{12})^T$

Мөрдлөгөө

Зөвхөн X_1 ба X_2 хамааралгүй үед л $\Sigma_{12} = 0$ байна.

Мөрдлөгөө

$X \sim N_p(\mu, \Sigma)$ бол зөвхөн $A\Sigma B^T = 0$ үед л AX ба BX санамсаргүй хувьсагчид хамааралгүй байна.

Хэсэг 2

Нөхцөлт тархалт

Нөхцөлт тархалт

Теорем

$X_1 = x_1$ үеийн X_2 санамсаргүй векторын нөхцөлт тархалт нь

$$(X_2 | X_1 = x_1) \sim N_{p-r}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

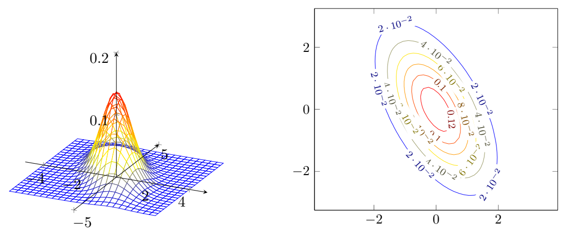
буюу $\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ дундаж болон $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ ковариацын матрицтай хэвийн тархалт байна.

Баталгаа

$X_2 = X_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}X_1$ тул $X_1 = x_1$ үед $X_2 = X_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1$ байна. Иймд $X_{2.1} \sim N_{p-r}(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$ байдаг тул X_2 нь $EX_2 = EX_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1 = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ дундаж $\operatorname{cov}(X_{2.1} + \Sigma_{21}\Sigma_{11}^{-1}x_1) = \operatorname{cov}(X_{2.1}) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ ковариацын матрицтай олон хэмжээст хэвийн тархалттай болно. □

Жишээ

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



нягтын функцийн график

нягтын функцийн түвшний шугам

Зураг: X_1 ба X_2 санамсаргүй хувьсагчдын хамтын тархалт

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

$$E(X_2|X_1 = x_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) = -0.8 \cdot 1 \cdot x_1 = -0.8x_1$$

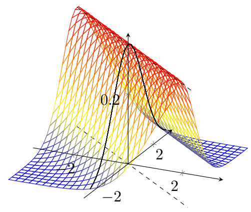
нөхцөлт ковариац нь

$$\operatorname{cov}(X_2|X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 2 - (0.8)^2 = 1.36$$

гэж олдоно. Иймд $X_1 = x_1$ үеийн X_2 хувьсагчийн нөхцөлт тархалт нь $N_1(-0.8x_1, 1.36)$ буюу

$$f(x_2|x_1) = \frac{1}{\sqrt{2\pi 1.36}} \exp\left\{-\frac{(x_2 + 0.8x_1)^2}{2(1.36)}\right\}$$

гэж олдоно.



Зураг: $f(x_2|x_1) = \frac{1}{\sqrt{2\pi 1.36}} \exp\left\{-\frac{(x_2 + 0.8x_1)^2}{2(1.36)}\right\}$ нөхцөлт нягтын муруй

Нөхцөлт тархалт болон тухайн тархалтаар хамтын тархалт байгуулах

Теорем

$X_1 \sim N_r(\mu_1, \Sigma_{11})$ ба $(X_2|X_1 = x_1) \sim N_{p-r}(Ax_1 + b, \Psi)$ бол

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

болно. Энд

$$\mu = \begin{pmatrix} \mu_1 \\ A\mu_1 + b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{11}A^T \\ A\Sigma_{11} & \Psi + A\Sigma_{11}A^T \end{pmatrix}$$

байна.

Жишээ

$X_1 \sim N_1(0, 1)$ ба

$(X_2|X_1 = x_1) \sim N_2\left(\begin{pmatrix} 2x_1 \\ x_1 + 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ бол $X = (X_1, X_2, X_3)$ хувьсагчдын хамтын тархалтыг ол.

Энэ тохиолдолд

$$Ax_1 + b = \begin{pmatrix} 2x_1 \\ x_1 + 1 \end{pmatrix} \Longrightarrow A = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ ба } b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

бас $\Psi = I_2$ тул

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

гэж олдоно.

Хэсэг 3

Нөхцөлт дөхөлт

Нөхцөлт дөхөлт

$X = (X_1, X_2)$, $X_1 \in \mathbb{R}^r$, $X_2 \in \mathbb{R}^{p-r}$ байг.

$$X_2 = h(X_1) + U$$

Теорем

$E(X_2|X_1)$ нь X_1 хувьсагчийн тусламжтайгаар X_2 уруу дөхөх $h(X_1) : \mathbb{R}^r \to \mathbb{R}^{p-r}$ функцүүд дундаас хамгийн бага дундаж квадрат алдаатай нь юм.

Дундаж квадрат алдаа $E\left\{(X_2 - h(X_1))^T(X_2 - h(X_1))\right\}$ гэж тодорхойлдог.

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

Чанар

$EU = 0$

Регрессийн шугаман загвар

X_1 болон X_2 санамсаргүй векторууд хамтдаа олон хэмжээст хэвийн тархалттай бол

$$\begin{aligned} X_2 &= E(X_2|X_1) + U \\ &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1) + U \\ &= \beta_0 + BX_1 + U \end{aligned}$$

болно. Энд $B = \Sigma_{21}\Sigma_{11}^{-1}$, $\beta_0 = \mu_2 - B\mu_1$,

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

байна.

$$X_2 = \beta_0 + BX_1 + U$$

шугаман загвар нь тухайлбал $r = p - 1$ тохиолдолд

$$X_2 = \beta_0 + \beta^T X_1 + U \qquad \text{энд } \beta^T = B_{(1\times r)} \text{ мөр вектор}$$

буюу

$$X_p = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r + U$$

хэлбэртэй болно.

Жишээ

$X = (X_1, X_2, X_3)$ буюу $p = 3$ үед $r = 3 - 1 = 2$ тул $X_1 = (X_1, X_2)$ ба $X_2 = (X_3)$ болж улмаар

$$X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

загвар бичигдэнэ.

Тодорхойлолт (Детерминацийн коэффициент)

$$\underbrace{\mathrm{cov}(X_2)}_{\text{нийт ковариац}} = \underbrace{\mathrm{cov}(\beta_0 + BX_1)}_{\text{тайлбарлагдах ковариац}} + \underbrace{\mathrm{cov}(U)}_{\text{үл тайлбарлагдах ковариац}}$$

$$\rho^2 = \frac{\text{тайлбарлагдах ковариац}}{\text{нийт ковариац}}$$

Детерминацийн коэффициентыг $r = p - 1$ тохиолдолд олж гаргавал

$$\rho^2 = \frac{\mathrm{cov}(\beta_0 + \beta^T X_1)}{\mathrm{cov}(X_2)} = \frac{\beta^T \mathrm{cov}(X_1)\beta}{\mathrm{cov}(X_2)} = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}}$$

болно. Энд $\sigma_{22} = \Sigma_{22} = DX_2$ нь скаляр, $\sigma_{21} = \Sigma_{21}$ нь r ширхэг компоненттой мөр вектор, $\sigma_{12} = \Sigma_{12}$ нь r ширхэг компоненттой багана вектор байна.

Жишээ

$\Sigma = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix}$ бол $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ шугаман загварын детерминацийн коэффициентыг ол.

$r = 2$ тул

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

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

$$\rho^2 = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}} = 0.5$$

үр дүнд хүрнэ.


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

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

# Linear model
fit <- lm(formula = V3 ~ ., data = X)

# Coefficient of determination
summary(fit)$r.squared
```


Регрессийн шугаман загварын тайлбарлах хувьсагчдын өмнөх коэффициент сая үзсэнчлэн ерөнхий тохиолдолд

$$B = \Sigma_{21}\Sigma_{11}^{-1}$$

харин $r = p - 1$ тухайн тохиолдолд

$$\beta^T = \sigma_{21}\Sigma_{11}^{-1}$$

байна.

Жишээ

Өмнөх жишээний хувьд β_1 болон β_2 коэффициентуудыг олъё.

$\sigma_{21} = \begin{pmatrix} 1 & 2 \end{pmatrix}$, $\Sigma_{11}^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}$ тул дараах хариу олдоно.

$$\beta^T = \sigma_{21}\Sigma_{11}^{-1} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

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


Хэсэг 4

Параметрийн үнэлэлт

Параметрийн үнэлэлт

Томъёо

- $\hat{\mu} = \overline{X}$ түүврийн дундаж
- $\hat{\Sigma} = S$ түүврийн ковариацын матриц

μ параметрийн үнэлэлтийн гаргалгаа үзүүлье.

- Ашиглах арга: Хамгийн их үнэний хувь бүхий арга
- үл мэдэгдэх параметр: $\theta = (\mu, \Sigma)$
- үл мэдэгдэх параметрийн тоо: $p + \left(p + \dfrac{p^2 - p}{2}\right) = p + \dfrac{1}{2}p(p+1)$
- түүвэр: $\mathcal{X} = (x_1, \dots, x_n) = \begin{pmatrix} x_1^1 & \dots & x_1^n \\ \vdots & \ddots & \vdots \\ x_p^1 & \dots & x_p^n \end{pmatrix}$

Үнэний хувь бүхий функц

$$\begin{aligned} L(\mathcal{X}, \theta) &= f(x_1, \theta) \cdot \ldots \cdot f(x_n, \theta) \\ &= |2\pi\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right\} \end{aligned}$$

Логарифм үнэний хувь бүхий функц

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

Үүнийг максимумчлахын тулд онолын дундаж оролцсон, сөрөг утгатай хоёр дахь гишүүнийг минимумчлах шаардлагатай.

Квадратлаг хэлбэрийг минимумчлах зорилгоор хувиргавал

$$\begin{aligned} &\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \\ &= \sum_{i=1}^{n}(x_i - \overline{x})^T \Sigma^{-1}(x_i - \overline{x})^{2} + n(\overline{x} - \mu)^T \Sigma^{-1}(\overline{x} - \mu) \\ &= \sum_{i=1}^{n} \mathrm{tr}\left\{(x_i - \overline{x})^T \Sigma^{-1}(x_i - \overline{x})\right\}^{3} + n(\overline{x} - \mu)^T \Sigma^{-1}(\overline{x} - \mu) \\ &= \mathrm{tr}\left\{\Sigma^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^T(x_i - \overline{x})\right\} + n(\overline{x} - \mu)^T \Sigma^{-1}(\overline{x} - \mu) \\ &= \mathrm{tr}\{\Sigma^{-1} nS\} + n(\overline{x} - \mu)^T \Sigma^{-1}(\overline{x} - \mu) \end{aligned}$$

μ онолын дунджийг $\overline{x}$ түүврийн дунджаар солиход хүрч байна.

[2] скаляр
[3] $\mathrm{tr}(ABC) = \mathrm{tr}(BCA) = \mathrm{tr}(CAB)$

Хувиргалтын дараах логарифм үнэний хувь бүхий функц

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

Үүнийг μ параметрээр максимумчлах гэдгээс

$$\hat{\mu} = \overline{x}$$

гэсэн үр дүнд хүрнэ.

Лекц V

Олон хэмжээст хэвийн тархалт III

Олон хэмжээст хэвийн тархалт III сэдвийн агуулга

1 Үнэний хувь бүхий харьцаат шинжүүр

2 Параметрийн таамаглалууд

3 Дунджийн итгэх муж

Хэсэг 1

Үнэний хувь бүхий харьцаат шинжүүр

Үнэний хувь бүхий харьцаат шинжүүр

Таамаглал
$f_X(x)$ нягттай X санамсаргүй векторын тархалтын үл мэдэгдэх θ параметрийн тухайн

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

таамаглал авч үзье.
Энд H_1 буюу Θ_1 ерөнхий тохиолдлыг харин H_0 буюу Θ_0 тухайн (тусгай) тохиолдлыг илэрхийлнэ. Өөрөөр хэлбэл $\Theta_0 \subset \Theta_1$ байна.

Жишээ

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \text{ дээр нөхцөл, хязгаарлалт тавигдаагүй}$$

Өөрөөр хэлбэл $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \mathbb{R}^p$ байна.

Шинжүүрийн статистик
Үл мэдэгдэх θ параметр бүхий тархалттай X санамсаргүй векторын эх олонлогоос $\mathcal{X}$ түүвэр авсан гэе.

$$\lambda(\mathcal{X}) = \frac{L_0}{L_1} = \frac{\max_{\theta\in\Theta_0} L(\mathcal{X},\theta)}{\max_{\theta\in\Theta_1} L(\mathcal{X},\theta)}$$

Энд L нь үнэний хувь бүхий функц юм.

Шинжүүрийн няцаах муж
H_0 үнэн үед L_0 харин эсрэг тохиолдолд L_1 давамгайлах тул шинжүүрийн няцаах муж

$$\{\mathcal{X} : \lambda(\mathcal{X}) < c\}$$

улмаар

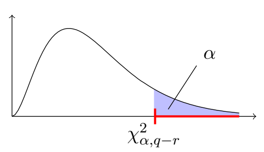
$$\{\mathcal{X} : -2\ln\lambda(\mathcal{X}) > t\}$$

хэлбэртэй болно.

Шинжүүрийн асимптот няцаах утга ба шинжүүрийн асимптот няцаах муж
α итгэх түвшинтэй шинжүүрийн асимптот няцаах муж Wilks-ийн теоремоор

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

болно. Энд r болон q нь харгалзан тэг болон өрсөлдөгч таамаглал буюу тусгай болон ерөнхий тохиолдлууд дахь үл мэдэгдэх параметрүүдийн тоо юм.



Зураг: Үнэний хувь бүхий харьцаат шинжүүрийн асимптот няцаах муж

Хэсэг 2

Параметрийн таамаглалууд

Таамаглал 1: 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

$$\ln L_0 = -\frac{1}{2} \ln |2\pi\Sigma| - \frac{n}{2} \operatorname{tr}\{\Sigma^{-1} S\} - \frac{n}{2} (\overline{x} - \mu_0)^T \Sigma^{-1} (\overline{x} - \mu_0)$$

$$\ln L_1 = -\frac{1}{2} \ln |2\pi\Sigma| - \frac{n}{2} \operatorname{tr}\{\Sigma^{-1} S\}$$

тул

$$-2 \ln \lambda = 2(\ln L_1 - \ln L_0) = n(\overline{x} - \mu_0)^T \Sigma^{-1} (\overline{x} - \mu_0)$$

Шинжүүрийн статистикийн асимптот хи-квадрат тархалтын чөлөөний зэрэг

$$q - r = p - 0 = p$$

Таамаглал 2: 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

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

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

R програмын `datasets` багцын `iris3` массивт хадгалаатай байдаг цахилдаг цэцгийн дэлбээ болон аяганы хэмжээний өгөгдөл авч үзье. Түүний Setosa төрөл зүйлийн хувьд цэцгийн дэлбээ болон аяганы 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)
# test statistic
as.numeric((n-p)/p * n/(n-1) * t(m - mu_0) %*% solve(S)
   %*% (m - mu_0))
# critical value
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)
```


Таамаглал 3: Ковариацын 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

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)
# test statistic
n * psych::tr(solve(Sigma_0) %*% S) - n *
   log(det(solve(Sigma_0) %*% S)) - n * p
# critical value
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.

Хэсэг 3

Дунджийн итгэх муж

Олон хэмжээст хэвийн тархалтын дундджийн итгэх муж

Таамаглал 2 буюу ковариац нь мэдэгдэхгүй байх үед дундджийн тухай таамаглал шалгах F шинжүүрээс олон хэмжээст хэвийн тархалтын дундаж утгын векторын итгэх мужийн дараах хэлбэртэй томъёо гарна.

Томъёо (Хэвийн тархалтын дундджийн $1-\alpha$ 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

Өмнөх жишээнд авч үзэж байсан **iris3** массиваас ялган авсан

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

өгөгдөл ашиглаж, цахилдагийн Setosa төрөл зүйлийн цэцгийн 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 дундджийн 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.

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.

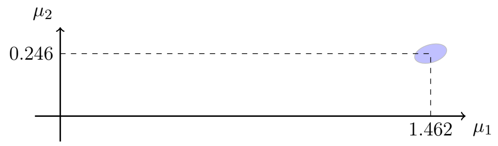
Өмнөх илэрхийлэл дэх статистикуудыг дараах байдлаар олно.

```
mean(X$`Petal L.`)   # x̄1 = 1.462
mean(X$`Petal W.`)   # x̄2 = 0.246
sd(X$`Petal L.`)     # σ1 ≈ 0.174
sd(X$`Petal W.`)     # σ2 ≈ 0.105
cor(x = X$`Petal L.`, y = X$`Petal W.`)   # ρ ≈ 0.332
```

Эдгээрийг орлуулбал дундджийн итгэх мужийн

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

илэрхийлэл олдох бөгөөд дор зурж дүрслэв.



Лекц VI

Олон хэмжээст хэвийн тархалт IV

Олон хэмжээст хэвийн тархалт IV сэдвийн агуулга

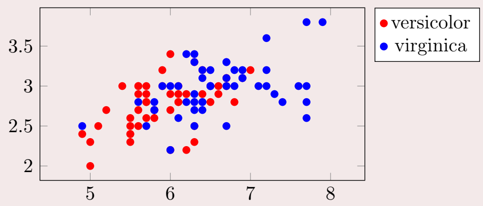
1 Дундаж утгын векторуудыг жиших
2 Ковариацын матрицуудыг жиших
3 Олон хэмжээст дисперсийн шинжилгээ

Хэсэг 1

Дундаж утгын векторуудыг жиших

Таамаглал 7: Дунджуудын зөрүүний тухай таамаглал

Жишээ



Зураг: Цахилдаг цэцгийн versicolor болон virginica төрөл зүйлүүдийн дэлбээний урт болон өргөн

Дэлбээний урт болон өргөний дундаж хэмжээ тус хоёр төрөл зүйлийн хувьд ялгаатай гэж батлах зорилго тавья.

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 ковариацын матрицуудын жинлэсэн дундаж буюу нийт түүврийн ковариацын 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}
# test statistic
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))
# critical value
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.

Таамаглал 9: Хоёр дундаж тэнцүү байх тухай таамаглал

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

Таамаглал

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

Шинжүүрийн статистик, түүний асимптот тархалт

$$(\overline{x}_1 - \overline{x}_2)^T \left(\frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\overline{x}_1 - \overline{x}_2) \sim \chi^2_p$$

Хэсэг 2

Ковариацын матрицуудыг жиших

Таамаглал 8: Ковариацууд 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 түүврүүд авсан гэе.

Таамаглал

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

Жишээ

Өмнөх жишээнд дунджууд тэнцүү гэсэн таамаглал шалгахад ашигласан шинжүүр нь бүлгүүдийн ковариацын матрицуудыг тэнцүү гэсэн нөхцөлтэй байсан бөгөөд бид тэрхүү нөхцөлийг шалгаж нягталгүйгээр 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 $u = \left(\sum_{i=1}^{k}\dfrac{1}{n_i-1} - \dfrac{1}{\sum_{i=1}^{k}(n_i-1)}\right)\dfrac{2p^2+3p-1}{6(p+1)(k-1)}$ байна.

Ковариацууд тэнцүү байх тухай таамаглалыг biotools багцын boxM() функц ашиглаж шалгах боломжтой.

```
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 ковариацын матрицуудыг ялгаатай гэх үндэслэлгүй ажээ.

Хэсэг 3

Олон хэмжээст дисперсийн шинжилгээ

Таамаглал 10: Дунджууд тэнцүү байх тухай таамаглал буюу Олон хэмжээст дисперсийн шинжилгээ (MANOVA)

Нэг ижил ковариацын 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 ковариацын матрицыг мэдэгдэхгүй гэж тооцно.

Шинжүүрийн статистик

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

Энд

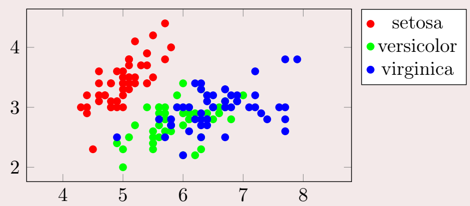
$$W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)^T \quad B = \sum_{i=1}^{k}(\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})^T$$

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



Зураг: Цахилдаг цэцгийн setosa, versicolor болон virginica төрөл зүйлүүдийн дэлбээний урт болон өргөн

Дэлбээний урт болон өргөний дундаж хэмжээ тус гурван төрөл зүйлийн хувьд ялгаатай болохыг батлах зорилго тавья.

Өгөгдөл бэлдэх

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

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
```


Лекц VII

Гол хэсгийн шинжилгээ

Гол хэсгийн шинжилгээ сэдвийн агуулга

1 Танилцуулга

2 Стандартчилсан шугаман эвлүүлэг

3 Гол хэсгүүдийг тайлбарлах нь

Хэсэг 1

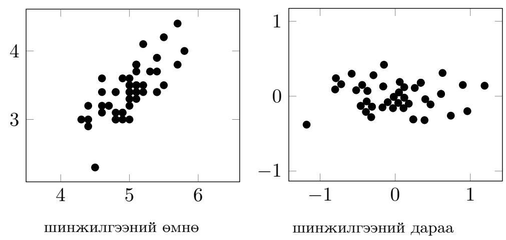
Танилцуулга

Гол хэсгийн шинжилгээ

Гол хэсгийн шинжилгээ буюу Principal Component Analysis (PCA) нь санамсаргүй векторын ковариацын бүтцийг задлан ялгадаг статистикийн арга техник юм. Тодруулбал энэ шинжилгээний тусламжтай өгөгдлийн хувьсан өөрчлөгдөж буй гол чиглэлүүдийг ялгаж салгадаг. Тийнхүү дисперс буюу хэлбэлзэл ихтэй тэдгээр чиглэлийн дагуу авсан хамааралгүй хувьсагчдын тусламжтай анх авсан хувьсагчдыг илэрхийлж болно. Тэдгээр шинэ хувьсагчдыг дисперсээр нь эрэмбэлж хамгийн бага дисперстэй буюу хамгийн бага мэдээлэл агуулж буй хувьсагчдыг орхивол хэмжээс бууруулсан явдал болно. Эсрэгээрээ хамгийн их дисперстэй цөөн хэдэн хувьсагч анхны өгөгдөлд байсан мэдээллийн дийлэнх хувийг агуулна.

Гол хэсгийн 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 зорилгоор өргөн хэрэглэж байна.


```r
X2 <- subset("x" = datasets::iris, "subset" = Species ==
    "setosa", "select" = c("Sepal.Length", "Sepal.Width"))
plot(X2, asp = 1)
pca <- princomp(x = X2, fix_sign = FALSE)
plot(x = pca$scores[,1:2], asp = 1)
```



Зураг: Хоёр хэмжээст X2 өгөгдөл дээрх гол хэсгийн шинжилгээ


```r
X4 <- subset("x" = datasets::iris, "subset" = Species ==
    "setosa", "select" = c("Sepal.Length", "Sepal.Width",
    "Petal.Length", "Petal.Width"))
```

Өгөгдөл

```r
head(X4)
```

Үр дүн

```
Sepal.Length Sepal.Width Petal.Length Petal.Width
         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

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

Үр дүн

```
Sepal.Length Sepal.Width Petal.Length Petal.Width
  0.12424898  0.14368980   0.03015918  0.01110612
```

Шинжилгээ

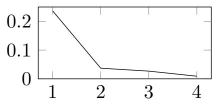
```r
pca <- princomp(x = X4)
```

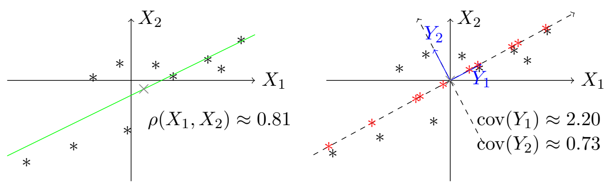
Дисперс

```r
apply(X = pca$scores,
    MARGIN = 2, FUN = var)
```

Үр дүн

```
     Comp.1      Comp.2      Comp.3      Comp.4
0.236455690 0.036918732 0.026796399 0.009033261
```





(a) Өгөгдөл (b) Үр дүн

Зураг: Гол хэсгийн шинжилгээ

- Цэнхэр – гол хэсгийн шинжилгээний үр дүн
- Ногоон – регрессийн шулуун $X_2 = -0.42 + 0.49X_1$
- Улаан – нэг дүгээр гол хэсэг буюу Y_1 тэнхлэг дээрх проекц
- Саарал – төв буюу $\overline{X} \approx (0.36, -0.24)^T$

Хэсэг 2

Стандартчилсан шугаман эвлүүлэг

Тархалтын төв буюу дунджийн тухай

Гол хэсгийн шинжилгээнд

$$EX = 0$$

нөхцөл биелдэг эсвэл

$$X := X - \mu$$

гэж хувиргасан гэж тооцдог.

Хамгийн сайн проекц

Проекцын чиглэл

X санамсаргүй векторын утгуудыг аль чиглэлд проекцлох вэ? Өөрөөр хэлбэл

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

хувиргалтын $\delta = (\delta_1, \ldots, \delta_p)^T$ векторыг хэрхэн сонгож авах вэ?

- δ_i нь X_i хувьсагчид ногдох жин
- $||\delta|| = 1$ буюу $\sum\limits_{i=1}^{p} \delta_i^2 = 1$

Зорилго

$$\operatorname{cov}(Y) = \operatorname{cov}(\delta^T X) \longrightarrow \max$$

Нийт дисперс, түүний задаргаа

Тодорхойлолт (Өргөтгөсөн дисперс)

$$|\Sigma| = |\operatorname{cov}(X)|$$

буюу X хувьсагчийн ковариацын матрицын тодорхойлогчийг X хувьсагчийн *өргөтгөсөн дисперс* гэе.

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

$\operatorname{cov}(\delta^T X) \longrightarrow \max$ бодлого

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

Бодлого

$\delta^T \delta = 1$ нөхцөлд

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

максимум олох

Бодолт

- Лагранжийн функц
$$\mathrm{L}(\delta, \lambda) = \delta^T \Sigma \delta - \lambda(\delta^T \delta - 1)$$
$$\frac{\partial\,\mathrm{L}}{\partial \delta} = 2\Sigma\delta - 2\lambda\delta = 0$$
- 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 ковариацын матрицын хувийн утга юм.

Проекцын вектор

$||\delta|| = 1$ болон $\max\limits_{\delta} \delta^T \Sigma \delta = \lambda_1$ 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**
$$Y_1 = \gamma_1^T X$$
- **Шинэ хувьсагчийн дисперс**
$$\operatorname{cov}(Y_1) = \lambda_1$$

Дараагийн гол хэсгүүд

Нэмэлт нөхцөл

$$\operatorname{cov}(Y_1, Y_2) = 0$$

$$\begin{aligned}\operatorname{cov}(Y_2, Y_1) &= \operatorname{cov}(\delta^T X, \gamma_1^T X) = \delta^T \operatorname{cov}(X) \gamma_1 = \delta^T \Sigma \gamma_1 \\ &= \lambda_1 \delta^T \gamma_1 = \lambda_1 \operatorname{cov}(\delta, \gamma_1) = 0\end{aligned}$$

$\lambda_1 > 0$ тул уг нэмэлт нөхцөл $\operatorname{cov}(\delta, \gamma_1) = 0$ болно.

Нэмэлт нөхцлийн сүүлийн хэлбэр

$$\operatorname{cov}(\delta, \gamma_1) = \delta^T \gamma_1 = 0$$

Бодлого: Дараагийн гол хэсэг

$\delta^T \delta = 1$ ба $\delta^T \gamma_1 = 0$ нөхцөлд $\max\limits_{\delta} \delta^T \Sigma \delta = ?$

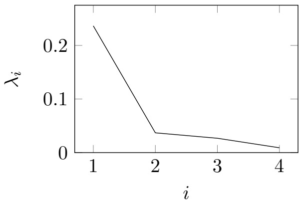
$$\mathrm{L}(\delta, \alpha, \beta) = \delta^T \Sigma \delta - \alpha(\delta^T \delta - 1) - \beta(\delta^T \gamma_1)$$
$$\frac{\partial\,\mathrm{L}}{\partial \delta} = 2\Sigma\delta - 2\alpha\delta - \beta\gamma_1 = 0$$
$$\delta^T \Sigma \delta = \alpha$$

хувийн векторын тодорхойлолт ёсоор δ максимумчлагч нь Σ матрицын α хувийн утганд харгалзах хувийн вектор болох бөгөөд λ_1 хувийн утгыг нэгэнт авсан тул одоо $\alpha = \lambda_2$ гэж авна.

Гол хэсгүүд, тэдгээрийн дисперс

$$Y_i = \gamma_i^T X, \qquad \operatorname{cov}(Y_i) = \lambda_i$$

энд, $i = 1, \dots, p$ бас $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$



Зураг: X4 өгөгдөл дээр хийсэн шинжилгээнээс гарсан хувийн утгууд

Бодолтын нэгдсэн үр дүн буюу эцсийн хариу

Томьёо

$EX = \mu$ ба $\operatorname{cov}(X) = \Sigma = \Gamma\Lambda\Gamma^T$ бол

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

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

болно.

Зарим чанар

Чанар

1. $EY_i = 0$
2. $\operatorname{cov}(Y_i) = \lambda_i$
3. $\operatorname{cov}(Y_i, Y_j) = 0,\ i \neq j$
4. $\operatorname{cov}(Y_1) \geq \dots \geq \operatorname{cov}(Y_p) \geq 0$
5. $\sum_{i=1}^{p} \operatorname{cov}(Y_i) = \operatorname{tr}(\Sigma)$
6. $\prod_{i=1}^{p} \operatorname{cov}(Y_i) = |\Sigma|$

Баталгаа (6 дугаар чанар)

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

$\square$

Хэсэг 3

Гол хэсгүүдийг тайлбарлах нь

Гол хэсгүүдийг тайлбарлах нь

- X_j хувьсагчдын өмнөх коэффициент буюу ”жин”

$$Y_i = \gamma_{i1}X_1 + \dots + \gamma_{ip}X_p$$

- Гол хэсэг ба санамсаргүй хувьсагч хоорондын ковариац

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

- Y_i гол хэсэг ба X_j хувьсагч хоорондын корреляц

$$\rho_{X_iY_j} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{X_iX_i}\lambda_j)^{1/2}} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_iX_i}}\right)^{1/2}$$

Чанар

$$\rho^2_{X_iY_1} + \dots + \rho^2_{X_iY_p} = 1$$

Баталгаа

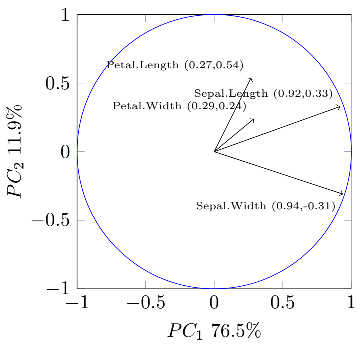
$\Sigma = \Gamma^T\Lambda\Gamma$ болохыг анхаарвал

$$\rho^2_{X_iY_1} + \dots + \rho^2_{X_iY_p} = \sum_{j=1}^{p} \rho^2_{X_iY_j} = \frac{\sum_{j=1}^{p} \lambda_j\gamma^2_{ij}}{\sigma_{X_iX_i}} = \frac{\gamma_i^T\Lambda\gamma_i}{\sigma_{X_iX_i}} = \frac{\sigma_{X_iX_i}}{\sigma_{X_iX_i}} = 1$$

болно. $\square$

Мөрдлөгөө

Дурын $q \leq p$ бүрийн хувьд $\rho^2_{X_iY_1} + \dots + \rho^2_{X_iY_q} \leq 1$ байна.



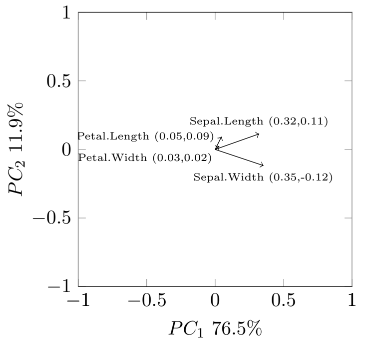
Зураг: X4 өгөгдөл дээр хийсэн шинжилгээнд харгалзах $\rho(X, Y)$ корреляцаар байгуулсан диаграмм

Гол хэсгийн шинжилгээ ба сингуляр утгын задаргаа

Практикт Σ ковариацын матрицыг өмнө үзсэнчлэн түүврийн ковариацын матрицаар үнэлнэ. $EX = 0$ нөхцөл тавьсан тул $\mathcal{X}$ өгөгдлийн матрицаар түүврийн ковариацын матрицыг $S = \frac{1}{n-1}\mathcal{X}^T\mathcal{X}$ байдлаар олж болно. Нөгөө талаас түрүүн үзсэнчлэн ковариацын матрицын хувийн утгын задаргаагаа ашиглавал

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

болно. Иймхүү тус шинжилгээ $\mathcal{X} = U\Lambda^{1/2}\Gamma^T$ сингуляр утгын задаргаатай холбогдоно. Эндээс санамсаргүй хувьсагчидтай холбогдох $\Lambda^{1/2}\Gamma^T$ хэсгийг нь авч санамсаргүй хувьсагчдын гол хэсгууд дээрх проекцыг тодорхойлж болно. Улмаар үүний квадрат буюу ”дисперс”-ийн гол хэсгийн дисперсэд эзлэх хувийг тооцох замаар гол хэсэг дээрх санамсаргүй хувьсагчийн нөлөө буюу оролцоог олдог.



Зураг: X4 өгөгдөл дэх хувьсагчдын гол хэсгүүд дээрх проекц

Гол хэсэг дээрх санамсаргүй хувьсагчдын оролцоо

Гол хэсэг дээрх санамсаргүй хувьсагчдын нөлөө буюу оролцоог дараах томьёогоор олно.

$$\frac{p_{ji}^2}{\lambda_i}\cdot 100\%$$

Энд p_{ji} нь X_j хувьсагчийн Y_i гол хэсэг дээрх проекц юм.

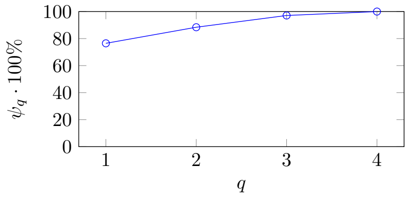
X4 өгөгдөл дэх хувьсагчдын гол хэсгүүд дээрх оролцоо

	Dim.1	Dim.2	Dim.3	Dim.4
Sepal.Length	44.7665911	35.746529	19.356724	0.13015586
Sepal.Width	53.8973034	38.523549	7.540926	0.03822131
Petal.Length	0.9320724	24.015448	69.297217	5.75526267
Petal.Width	0.4040331	1.714474	3.805133	94.07636016

Ялгаж авах гол хэсгийн тоо

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

эхний q гол хэсгээр "тайлбарлагдах" дисперсийн нийт дисперсэд эзлэх хувь



Зураг: X4 өгөгдөл дээр хийсэн шинжилгээнээс гарсан хувийн утгууд дээрх ψ_q харьцаа

Лекц VIII

Факторын шинжилгээ - I хэсэг

Факторын шинжилгээ - I хэсэг сэдвийн агуулга

1 Факторын шинжилгээний загвар

2 Шугаман хамааралтай хувьсагчид дээрх шинжилгээ

3 Шугаман хамааралгүй хувьсагчид дээрх шинжилгээ

4 Загварын онцлог

Хэсэг 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 ковариацын матриц бүхий

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

санамсаргүй вектор авч үзье.

Загвар

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

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

$X=LF+\mu$ 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.

Ковариацын 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

Ковариацын 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

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

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

Хэсэг 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

Гол хэсгийн шинжилгээ дэх $Y = \Gamma^T(X-\mu)$ хувиргалт

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

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

$$\begin{aligned}EF &= E(\Lambda_1^{-1/2} Y_1) \\ &= \Lambda_1^{-1/2} EY_1 \\ &= \Lambda_1^{-1/2} E(X_1 - \mu_1) \\ &= 0\end{aligned}$$

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

Хэсэг 3

Шугаман хамааралгүй хувьсагчид дээрх шинжилгээ

Шугаман хамааралгүй хувьсагчид дээрх шинжилгээ

Энэ тохиолдолд шууд орхиж болох хувьсагч байхгүй тул X санамсаргүй векторын ковариацын 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_j тухайн нэг хувьсагчийн хувьд

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

хэлбэрээр бичигдэнэ.

Ковариацын задаргаа

Нийт ковариацын 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

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

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.

Хэсэг 4

Загварын онцлог

Факторын шинжилгээний загвар масштабаас үл шалтгаалах нь

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

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

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 $\operatorname{cov}(Y) = \rho(Y)$ 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. $\rho(X, F) = L$ байсан тул дээрх эргүүлэлтийг тус корреляцыг хамгийн их байлгах гэх мэтчилэн ашигтай байдлаар сонговол зохимжтой. Тийнхүү эргүүлэлтийн тодорхойгүй байдал буюу факторын ачилтын олон утгат байдлаас зайлсхийхийн тулд

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

зэрэг нэмэлт нөхцөл тавьдаг.

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

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

$d < 0$ 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

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

$p = 3$ үед

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

$p = 4$ үед

$$d = \begin{cases} 2, & k = 1 \\ \le -1, & k \ge 2 \end{cases}$$

$p = 6$ үед

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

Лекц IX

Факторын шинжилгээ - II хэсэг

Факторын шинжилгээ - II хэсэг сэдвийн агуулга

1 Загварын үнэлгээ

2 Факторын үнэлгээ

3 Эргүүлэлт

Хэсэг 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 дараалаар хийнэ.

Сэргээн санах нь

Факторын шинжилгээний $X = LF + U + \mu$ загвар ашиглаж гаргасан ковариацын задаргааг шинжилгээний загвар гэж ойлгож болох бөгөөд нэг дүгээрт

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

задаргаа, хоёр дугаарт тус задаргаа дахь L ба Ψ үл мэдэгдэх параметрүүдийг үнэлэх бодолтын явцад гарч ирдэг

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

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 ковариацын задраанд тулгуурласан ойролцоо бодолт хийж шийд олно.

Сэргээн санах нь

Факторын ачилт L нь масштабаас үл шалтгаална.

Дээрх чанараас үүдэн загвар үнэлэхэд ковариацын матриц ба корреляцийн матрицын алийг нь ч ашиглаж болно. Ковариацын 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 ковариацын матриц болон түүврийн корреляцын матриц юм.

Жишээ

$p = 3$ 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 ажиглагалана уу. Бас $\hat{q}_2^2$ болон $\hat{q}_3^2$ ачилтуудыг ч дээрх байдлаар олж болно.

q_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

R программын **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)
# output
32
```

Эхний 6 мөрийг нь хэвлэх

```
head(X)
# output
                   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 $D = \operatorname{diag}(s_{X_1,X_1}, \ldots, s_{X_p,X_p}) = I$ болох тул

$$L^T D^{-1} L \text{ диагонал}$$

нэмэлт нөхцөл

$$L^T L \text{ диагонал}$$

хэлбэрт шилжинэ. Иймд L матрицын баганууд 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 хувь бүхий харьцаат шинжүүр зохиосон байдаг.
Гол хэсгийн арга Гол хэсгийн шинжилгээ дээр үзэж байсан ковариацын матрицын хувийн утгын задаргаанаас факторын ачилтыг үнэлдэг.

Гол хэсгийн аргаар загвар үнэлэх

1. **Түүврийн ковариацын матрицын хувийн утгын задаргаа**
$$S = \Gamma\Lambda\Gamma^T$$
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 $S - \hat{L}\hat{L}^T$ 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.

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

Иймхүү

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

үнэлэлт гарна.

Түүврийн ковариацын матриц

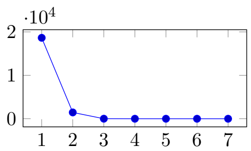
```
S <- cov(X)
```

Хувийн утга болон хувийн вектор

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

Хувийн утгууд

```
1.864032e+04 1.453913e+03 9.287215e+00 1.547302e+00
   3.767684e-01 9.632947e-02 8.008942e-02
```



$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_{12}} \cdot 100\% \approx 99.94\%$$

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,]    5.2043264  -0.3501269
[2,]   -1.6431818   0.1286879
[3,] -122.8214135 -16.6039392
[4,]  -59.3611358  34.3091285
[5,]    0.3631939   0.1487014
[6,]   -0.8518824  -0.1855300
[7,]    0.9108076  -0.9539552
```

Нийлэмж

Сэргээн санах нь

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

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

```
2.720760e+01 2.716607e+00 1.536079e+04 4.700861e+03
    1.540219e-01 7.601250e-01 1.739601e+00
```

Хөндлөнгийн хүчин зүйлсийн дисперсийн үнэлгээ

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

```
 mpg  cyl disp   hp drat   wt qsec
9.12 0.47 0.01 0.01 0.13 0.20 1.45
```


R програмын psych багц дахь principal() функц

Багц суулгах

```
install.packages("psych")
```

Шинжилгээ хийх

```
result <- psych::principal(r = X, nfactors = 2, covar =
    TRUE, rotate = "none")
```

Шинжилгээний үр дүнгийн тойм

```
print(result)
```

Тус функцийн утгаас өмнө бодож олсон чөлөөний зэрэг, түүврийн хэмжээ, хувийн утгууд, факторын ачилт, нийлэмж, хөндлөнгийн хүчин зүйлийн дисперс зэргийг дараах байдлаар гарган авч болно.

```
result$dof
result$n.obs
result$values
```

```
result$loadings
result$communality
result$uniquenesses
```


psych::principal() функцийн тоймлосон үр дүнгийн зарим хэсгийн тайлбар

- h2 нийлэмж буюу факторуудаар илэрхийлэгдэх дисперс h_i^2
- u2 хөндлөнгийн хүчин зүйлийн дисперс ψ_{ii}
- H2 нийлэмжийн хувьсагчийн дисперсэд эзлэх хувь h_i^2/s_{ii}
- U2 хөндлөнгийн хүчин зүйлийн дисперсийн хувьсагчийн дисперсэд эзлэх хувь ψ_{ii}/s_{ii}
- SS loadings хувийн утгууд
- Proportion Var хувийн утга тус бүрийн нийт хувийн утгын нийлбэрт эзлэх хувь
- Cumulative Var хувийн утгуудын хуримтлагдах нийлбэрийн нийт хувийн утгын нийлбэрт эзлэх хувь
- Proportion Explained хувийн утга тус бүрийн эхний k ширхэг хувийн утгын нийлбэрт эзлэх хувь
- Cumulative Proportion хувийн утгуудын хуримтлагдах нийлбэрийн эхний k ширхэг хувийн утгын нийлбэрт эзлэх хувь

Гол хэсгийн аргаар үнэлсэн загварын алдаа

S 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?

$$\|S - \hat{L}\hat{L}^T - \hat{\Psi}\|_F = \sqrt{\sum_{i,j}(S - \hat{L}\hat{L}^T - \hat{\Psi})_{ij}^2}$$

$$\leq \sqrt{\lambda_{k+1}^2 + \ldots + \lambda_p^2} = \|(\lambda_{k+1}, \ldots, \lambda_p)\|_F$$

Дээрх нормуудыг дараах байдлаар олж болно.

```
norm(result$residual, type = "F")
# output
1.817
norm(as.matrix(result$values[-{1:k}]), type = "F")
# output
9.423
```


Хэсэг 2

Факторын үнэлгээ

Факторын үнэлгээ

Факторын үнэлэлтийг *факторын оноо* гэдэг. $X - \mu$ болон F хамтдаа хэвийн тархалттай гэвэл олон хэмжээст хэвийн тархалт сэдэвт үзсэн ёсоор

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

ба $E(F|X = x) = L^T\Sigma^{-1}(X - \mu)$ байх тул

$$\hat{F} = L^T S^{-1}(X - \overline{X})$$

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

```
F <- t(t(L) %*% solve(S) %*% {t(X) - colMeans(X)})
```

psych::principal() функцийн хувьд факторын оноо нь түүний буцаах утгын scores элементэд хадгалагдаж байдаг.

Хэсэг 3

Эргүүлэлт

Эргүүлэлт

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

Сэргээн санах нь

$\rho(X, F) = L$

Эргүүлэлтийг $\rho(X, F) = L$ 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.

```
result <- psych::principal(r = X, nfactors = k, covar =
    TRUE, 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 хэсэг

Кластерын шинжилгээ - I хэсэг сэдвийн агуулга

1 Кластерын шинжилгээ

2 Кластер байгуулах алгоритмуудаас

3 Шатлах алгоритм

Хэсэг 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 бүлэглэл тогтоодог. Шинжилгээний зорилго бол өөр бүлгүүдэд байх юмс үзэгдлээс илүү адил төстэй юмс үзэгдлийг нэг бүлэгт бүлэглэх явдал юм. Тэрхүү бүлгийг *кластер* гэдэг. Үүний нөгөө талаас нь харвал ялгаатай юмс үзэгдлийг ангилан ялгах байдлаар кластер гэх бүлгүүдэд хуваарилж буй явдал юм.

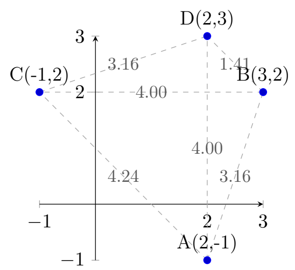
Жишээ

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

Хүснэгт: Кластерын шинжилгээнд ашиглах өгөгдөл

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

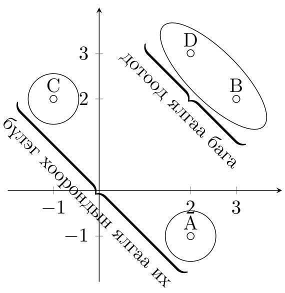
```
X <- matrix(data = c(2,-1,3,2,-1,2,2,3), nrow = 4, byrow =
    TRUE)
print(X)
```

Зураг: Объектууд буюу цэгүүд, тэдгээрийн хоорондох Евклидийн зай

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

Кластерын шинжилгээний үндсэн зарчим



Зураг: Кластер байгуулах үндсэн зарчим

Хэсэг 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.
- Бусад алгоритмууд

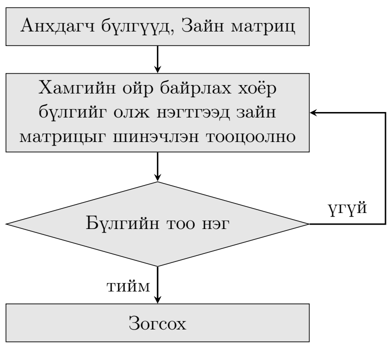
Хэсэг 3

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

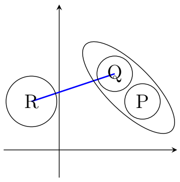
Томьёо ($P + Q$ шинэ бүлэг ба R бүлэг хоорондын зай)

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

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[1]	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	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

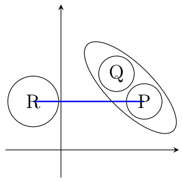
[1] жинлэсэн; n_P, n_Q, n_R харгалзан P, Q, R бүлэг дэх объектуудын тоо

Single linkage алгоритм



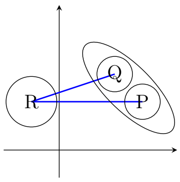
$$d(R, P+Q) = \frac{1}{2}\,d(R,P) + \frac{1}{2}\,d(R,Q) - \frac{1}{2}|\,d(R,P) - d(R,Q)| = \min\{d(R,P), d(R,Q)\}$$

Complete linkage алгоритм



$$d(R, P+Q) = \frac{1}{2}\,d(R,P) + \frac{1}{2}\,d(R,Q) + \frac{1}{2}|\,d(R,P) - d(R,Q)| = \max\{d(R,P), d(R,Q)\}$$

Average linkage алгоритм



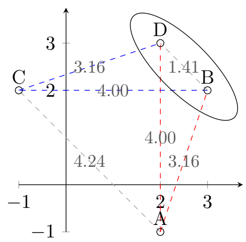
- жинлээгүй

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

- жинлэсэн

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

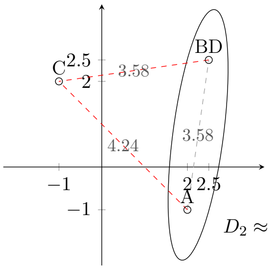
Жинлэсэн Average linkage алгоритм, I итерац



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

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

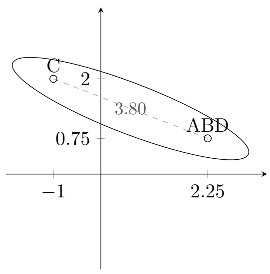
Жинлэсэн Average linkage алгоритм, II итерац



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

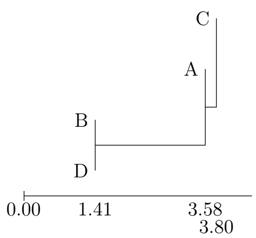
$$D_2 \approx \begin{pmatrix} 0 & \frac{1}{3}\cdot 4.24 + \frac{2}{3}\cdot 3.58 = \color{red}{3.80} \\ & 0 \end{pmatrix}$$

Жинлэсэн Average linkage алгоритм, III итерац



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

Жинлэсэн Average linkage алгоритмаар хийсэн кластерын шинжилгээний үр дүнг харуулсан дендрограмм

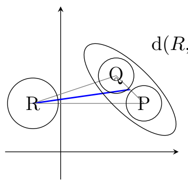



```
cl <- hclust(d, method = "average")
plot(cl)
cl$height
```


Centroid болон Median алгоритмууд

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

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



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

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

Ward алгоритм

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

Бүлгийн дотоод ялгаа дараах томьёогоор 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 тооцох аргачлал

- "single"
- "complete"
- "average"
- "median"
- "centroid"
- "ward.D"

Лекц XI

Кластерын шинжилгээ - II хэсэг

Кластерын шинжилгээ - II хэсэг сэдвийн агуулга

1 Хэсэгчлэх алгоритм

2 Кластерын төв

3 Өөр кластерт илүү ойр объект олох нь

4 Хувьсагчдын масштабын ялгаатай байдлын нөлөөг зайлуулах

Хэсэг 1

Хэсэгчлэх алгоритм

Кластерын шинжилгээний хэсэгчлэх алгоритм

Сэргээн санах нь

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.

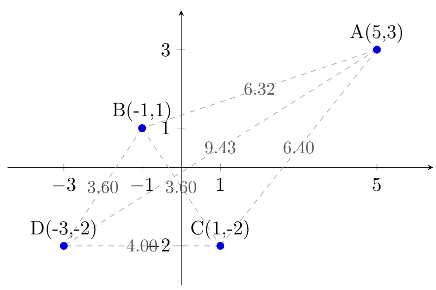
Жишээ

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

Хүснэгт: Кластерын шинжилгээнд ашиглах өгөгдөл

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)
```

Зураг: Объектууд буюу цэгүүд, тэдгээрийн хоорондох Евклидийн зай

Хэсэгчлэх алгоритмаар хийх кластерын шинжилгээний kmeans() функц

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

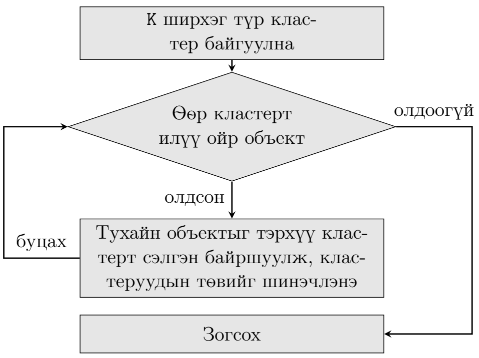
`x` өгөгдөл

`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



Хэсэг 2

Кластерын төв

Кластерын төв

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

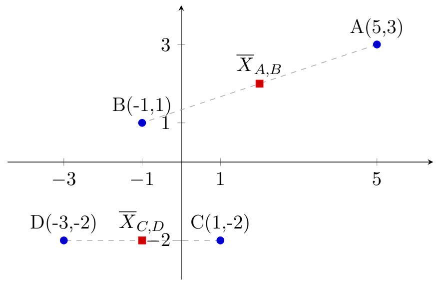
Томьёо (Кластерын төв шинэчлэх)

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

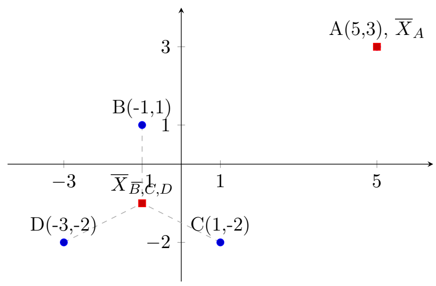
$$\overline{X}_1 = \frac{2\cdot(2)-(-1)}{2-1} = 5 \qquad \overline{X}_2 = \frac{2\cdot(2)-1}{2-1} = 3$$

- 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 $\{B, C, D\}$



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

Хэсэг 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.

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)

Энд d^2 нь зайн квадрат юм.

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

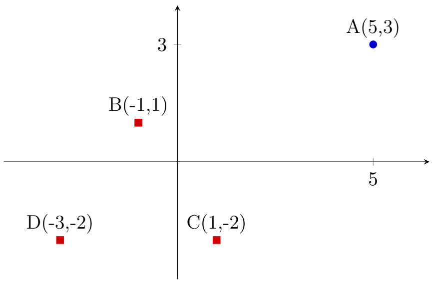
A болон B объект

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

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

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

$$SST = SSI + SSB$$

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

Хэсэг 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. Иймд 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 $A = \mathrm{diag}(s_{X_1X_1}^{-1}, \ldots, s_{X_pX_p}^{-1})$ буюу хувьсагчдын дундаж квадрат хазайлт ашиглаж болно. Үгтаа энэ нь хувьсагчдыг стандарт хувиргалтаар хувиргасантай адил болох тул тухайлбал R програм дээр өгөгдлийн матрицаа `X <- scale(X)` байдлаар хувиргана.

Лекц XII

Дискриминантын шинжилгээ

Дискриминантын шинжилгээ сэдвийн агуулга

1 Дискриминантын шинжилгээ

2 Ангиллын зарчим

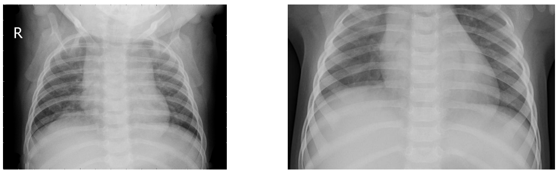
Хэсэг 1

Дискриминантын шинжилгээ

Дискриминантын шинжилгээний тусламжтай шийдэж болох асуудлын нэг жишээ

Жишээ

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: Цээжний рентген зураг

Дискриминантын шинжилгээг хэдийд хийж болох тухай

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

```
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

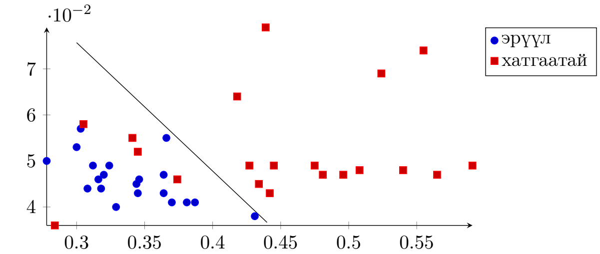
$$\text{зураг} \xrightarrow{\text{RGB}} \underbrace{\text{grayscale}}_{\text{матриц}} \to \text{SVD} \to \underbrace{\text{сингуляр буюу хувийн утгууд}}_{\text{вектор}}$$

 $\longrightarrow$ (0.565, 0.047, bacteria)

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

Хүснэгт: Бэлдсэн өгөгдлийн зарим мөр

Дискриминантын шинжилгээний үр дүн



Зураг: Дискриминантын шинжилгээний үр дүн

Санамж

Сегментчлэл, нормалчлал зэрэг үйлдэл нэмж хийвэл ангиллын алдаа улам буурч болох юм.

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 тогтоохоод 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.

Хэсэг 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

$$R_j = \{x : x \in \Pi_j\}$$

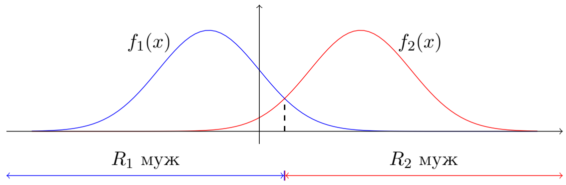
гэе.

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: 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

Санамсаргүй хувьсагчийн эх олонлогуудын тархалт $N_1(\mu_j, \sigma_j)$ 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

$$f_1(x) \geq f_2(x)$$

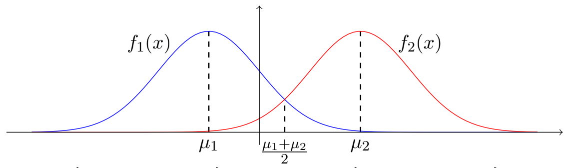
$$\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \geq \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}$$

$$x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) \leq 2\ln\frac{\sigma_2}{\sigma_1}$$

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.



$$R_1 = \left\{x : x \leq \frac{\mu_1 + \mu_2}{2}\right\} \text{ муж} \quad R_2 = \left\{x : x > \frac{\mu_1 + \mu_2}{2}\right\} \text{ муж}$$

Зураг: 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: 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

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:
   $$R_1 = \{x : \alpha^T(x - \mu) \geq 0\}$$
   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 ковариацыг адил тэнцүү гэж үзсэн. Иймд ковариацын матрицыг нийт түүврийн ковариацын матрицаар үнэлсэн.

$$\hat{\Sigma} = S \approx \begin{pmatrix} 0.00739 & 0.00022 \\ 0.00022 & 0.00008 \end{pmatrix}$$

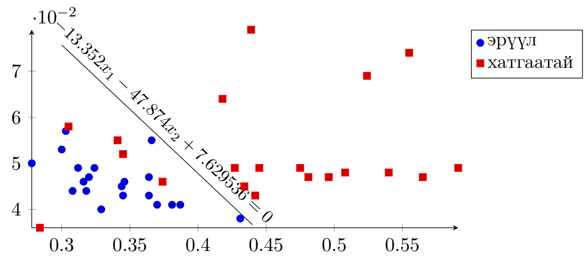
Харин эх олонлогуудын дунджийн үнэлэлт

$$\hat{\mu_1} \approx (0.340, 0.046) \text{ ба } \hat{\mu_2} \approx (0.449, 0.053)$$

гэж олдсон.

Ангиллын $R_1 = \{x : \alpha^T(x - \mu) \geq 0\}$ муж дараах хэлбэртэй болно.

$$\begin{aligned} \alpha^T(x - \mu) &\approx (-13.352, -47.874)\begin{pmatrix} x_1 - 0.395 \\ x_2 - 0.049 \end{pmatrix} \\ &\approx -13.352x_1 - 47.874x_2 + 7.629536 \geq 0 \end{aligned}$$



Зураг: Дискриминантын шинжилгээний үр дүн

Лекц XIII

Олон хэмжээст координатын шинжилгээ

Олон хэмжээст координатын шинжилгээ сэдвийн агуулга

1. Олон хэмжээст координатын шинжилгээний тухай ерөнхий ойлголт
2. Цэгийн координат сэргээх
3. Огторгуйн хэмжээс сонгох
4. Хувьсагчдын масштабын ялгаатай байдлын нөлөөг анхаарах

Хэсэг 1

Олон хэмжээст координатын шинжилгээний тухай ерөнхий ойлголт

Олон хэмжээст координатын шинжилгээ[1]

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.

[1] Multidimensional Scaling

Жишээ

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.

#	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: Аз жаргалын индекс, 2019 он, өгөгдлийн зарим хэсэг

Тэгвэл улс орнуудыг дээрх 6 хувьсагчийн утгуудаар ойролцоо эсвэл ялгаатай байдлыг харуулсан цэгэн диаграмм байгуул.

Сонгодог координатын шинжилгээний үндсэн санаа

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 тул хувьсагчид өөр хоорондоо хамааралтай ажээ. Иймд гол хэсгийн шинжилгээ хийж, нэр бүхий зургаан хувьсагчийг хамгийн сайн илэрхийлж чадах эхний хоёр гол хэсгээр улс орнуудын аз жаргалын ойролцоо эсвэл ялгаатай байдлыг харуулсан цэгэн диаграмм байгуулж болох юм. Өөрөөр хэлбэл түүврийн ковариацын матриц олж, түүний хувийн утгын задаргааг ашиглана.

R програмын cmdscale() функц

Зайн матриц

$$d = (d_{ij})_{i,j=1\ldots,n}$$

Энд $d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$ нь i болон j дугээр цэгүүдийн хоорондох Евклидийн зай юм.

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

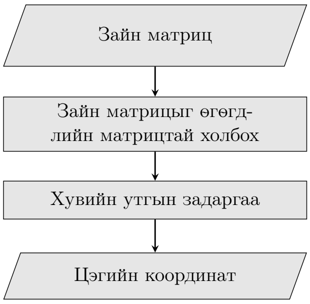
Шинжилгээ

```
result <- cmdscale(d = d, k = 2, eig = TRUE)
```


Хэсэг 2

Цэгийн координат сэргээх

Цэгийн координат сэргээх



Зураг: Цэгийн координат олох үйлдлийн дэс дараа

Зайн матрицыг өгөгдлийн матрицтай холбох

p огторгуйн хэмжээс $x_i = (x_{i1}, \dots, x_{ip})^T$ цэгийн координат
n цэгийн тоо $X = (x_1, \dots, x_n)^T$ өгөгдлийн матриц

Өгөгдлийн матриц дээр хувьсагчдын дундаж тэгтэй тэнцүү гэсэн нэмэлт нөхцөл тавъя.

$$
\begin{aligned}
d_{ij}^2 &= (x_{i1} - x_{j1})^2 + \dots + (x_{ip} - x_{jp})^2 \\
&= x_{i1}^2 + x_{j1}^2 - 2x_{i1}x_{j1} + \dots + x_{ip}^2 + x_{jp}^2 - 2x_{ip}x_{jp} \\
&= x_{i1}^2 + \dots + x_{ip}^2 + x_{j1}^2 + \dots + x_{jp}^2 - 2(x_{i1}x_{j1} + \dots + x_{ip}x_{jp}) \\
&= x_i^T x_i + x_j^T x_j - 2x_i x_j
\end{aligned}
$$

Энэхүү задаргаанд үндэслэн элементүүд нь

$$b_{ij} = x_i^T x_j = \sum_{k=1}^{p} x_{ik} x_{jk}$$

байх $B = (b_{ij})_{i,j=1,\dots,n}$ матриц зохиоё.

B матрицын чанар

Чанар

$$\sum_{i=1}^{n} b_{ij} = 0 \quad \forall j = 1, \dots, n$$

Баталгаа

Хувьсагчдын дундаж тэгтэй тэнцүү буюу $\sum_{k=1}^{p} x_{ik} = 0$ болохыг тооцвол

$$
\begin{aligned}
\sum_{i=1}^{n} b_{ij} &= \sum_{i=1}^{n} x_i^T x_j = (x_1 + \dots + x_n)^T x_j \\
&= (x_{1k} + \dots + x_{nk})_{k=1,\dots,p}^T x_j = 0
\end{aligned}
$$

болно. □

Зайн матрицаар B матрицыг илэрхийлэх

Зайн задаргааг ахин бичвэл $d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}$ болно. Одоо өмнөх чанарыг тооцон i, j бас i ба j индексүүдээр нийлбэрчилбэл

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 &= \frac{1}{n}\sum_{i=1}^{n} b_{ii} + b_{jj} \\
\frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 &= b_{ii} + \frac{1}{n}\sum_{j=1}^{n} b_{jj} \\
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 &= \frac{2}{n}\sum_{i=1}^{n} b_{ii}
\end{aligned}
$$

үр дүн гарна. Улмаар дээрх дөрвөн тэгшитгэлийг хамтатган бодвол

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2)$$

шийд олдоно. Энд $\cdot$ нь дунджийг илэрхийлнэ.

B матрицын бусад чанар

Чанар

1. $B = XX^T$
2. $\operatorname{rank}(B) = p$
3. B матриц эерэг хагас тодорхойлогдсон.

Баталгаа

1. B матрицын тодорхойлолтоос шууд харж болно.
2. Хувьсагчдыг шугаман хамааралгүй, тэг дээр бөхсөн тархалттай биш бас цэгүүдийн дор хаяж p ширхэг нь шугаман хамааралгүй гэж тооцсон үед $\operatorname{rank}(B) = \operatorname{rank}(XX^T) = \operatorname{rank}(X) = p$ болно.
3. Зайг Евклидийн нормоор авсан гэдгээс мөрдөн гарна.

□

B матрицын хувийн утгын задаргаа, цэгийн координат

Мөрдлөгөө

B матриц p ширхэг эерэг, $n - p$ ширхэг тэгтэй тэнцүү хувийн утгатай.

$$B = \Gamma\Lambda\Gamma^T$$

Энд $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ нь B матрицын хувийн утгуудаас тогтох матриц, $\Gamma = (\gamma_1, \dots, \gamma_p)$ нь хувийн векторуудаас тогтох матриц юм.

$B = XX^T$ чанарыг анхаарвал

$$B = \Gamma\Lambda\Gamma^T = \Gamma\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\Gamma^T = \underbrace{\Gamma\Lambda^{\frac{1}{2}}}_{X}\underbrace{\Lambda^{\frac{1}{2}}\Gamma^T}_{X^T} = XX^T$$

буюу

$$X = \Gamma\Lambda^{\frac{1}{2}}$$

гэж болно.

Хэсэг 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 тоо улмаар хувийн утгуудын

$$\psi_p = \frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

харьцаанд үндэслэж сонгодог. Энд $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ гэж тооцов.

Жишээний хувьд $p = 2$ үед

$$\psi_2 \approx \frac{41.845 + 4.541}{53.536} \approx 0.866$$

буюу нийт ковариацын ойролцоогоор 86.6 хувь нь 2 хэмжээст шинэ огторгуйд хадгалагдан үлдэж байна.

Хэсэг 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

Конжойнт шинжилгээ сэдвийн агуулга

1 Конжойнт шинжилгээ

2 Конжойнт шинжилгээний загвар

3 Загварын параметрийн үнэлгээ

4 Энгийн шугаман регрессийн загварт хувиргах нь

Хэсэг 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.

Бид энэхүү шинжилгээг эдийн засаг, маркетингийн талаас нь бус харин цэвэр статистикийн үүгээс авч үзнэ. Иймд тус шинжилгээг практикт хэрэглэх үеийн давуу эсвэл сул тал зэрэг зүйлсийг анхаарахгүйн дээр ханамж зэрэг эдийн засгийн нэр томъёо хэрэглэхгүй.

Аль хувилбар хамгийн сайн бэ?

Орц	Сав: жигнэмэг	Сав: хуванцар
аарц	?	?
шоколад	?	?
жимс	?	?

$X_1, \ldots, X_K$ бараа, үйлчилгээний үндсэн шинж чанар буюу 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 бүрийн түвшин буюу хувилбарын тоо

$M = L_1 \cdot \ldots \cdot L_K$ нийт хувилбарын тоо

Хэрэглэгч таашаалаараа хувилбаруудыг эрэмбэлсэн нь

Орц	Сав: жигнэмэг	Сав: хуванцар
аарц	2	1
шоколад	6	3
жимс	5	4

1 2 3 4 5 6

→ сайн

муу ←

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 6 \\ 3 \\ 5 \\ 4 \end{pmatrix}$$

Дээрх эрэмбэлэлт бол хувилбаруудын аль нь алинаасаа илүү болохыг заахаас хэтрэхгүй. Гэтэл Y_i нь угтаа илүү нарийвчлалтай тоо байна. Иймд $\hat{Y}_i$ үнэлэлт олох шаардлагатай.

R програмын conjoint багц ашиглаж $\hat{Y}_i$ үнэлэлт олох

```
conjoint::caTotalUtilities(
  y = c(2,1,6,3,5,4),
  x = expand.grid(
    packing = c("biscuit", "plastic"),
    ingredients = c("curd", "chocolate", "fruit")
  )
)
```

```
2.333 0.667 5.333 3.667 5.333 3.667
```

$$Y = \begin{pmatrix} 2 \\ 1 \\ 6 \\ 3 \\ 5 \\ 4 \end{pmatrix} \qquad \hat{Y} = \begin{pmatrix} 2.333 \\ 0.667 \\ 5.333 \\ 3.667 \\ 5.333 \\ 3.667 \end{pmatrix}$$

Хэсэг 2

Конжойнт шинжилгээний загвар

Конжойнт шинжилгээний загвар

$$Y_i = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \beta_{kl}\, \mathrm{I}(X_k = x_{kl}) + \mu + \epsilon_i, \quad \forall k : \sum_{l=1}^{L_k} \beta_{kl} = 0$$

- X_k факторууд $(k = 1, \ldots, K)$
- x_{kl} X_k факторын түвшингүүд $(l = 1, \ldots, L_k)$
- β_{kl} x_{kl} түвшингийн коэффициент $(l = 1, \ldots, L_k)$
- μ ерөнхий түвшин
- ϵ_i загварын алдаа

$$Y_1 = \beta_{11} + \beta_{21} + \mu \qquad Y_4 = \beta_{12} + \beta_{22} + \mu$$
$$Y_2 = \beta_{11} + \beta_{22} + \mu \qquad Y_5 = \beta_{13} + \beta_{21} + \mu$$
$$Y_3 = \beta_{12} + \beta_{21} + \mu \qquad Y_6 = \beta_{13} + \beta_{22} + \mu$$

Энд $\beta_{11} = \text{нөлөө}_{\text{орц, аарц}}$ ба $\beta_{21} = \text{нөлөө}_{\text{сав, жигнэмэг}}$ гэх мэтчилэн ойлгоно. Харин $\mu = \frac{1+2+3+4+5+6}{6} = 3.5$ байна.

Конжойнт шинжилгээний загварын зарим шинж чанар

1. Конжойнт шинжилгээний загвар нь нөхцөлт шугаман регрессийн загвар юм.
2. $\sum_{l=1}^{L_k} \beta_{kl} = 0$ нөхцөл нь тухайн нэг хувилбарт факторын бүх түвшинг тооцох нь эргээд тухайн факторын аль ч түвшинг тооцоогүйтэй адил болохыг илэрхийлнэ.

Хэсэг 3

Загварын параметрийн үнэлгээ

$n = 1$ хэрэглэгчээс авсан судалгааны мэдээлэлд үндэслэсэн загварын параметрийн үнэлгээ

X_1	X_2: 1	X_2: 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
$\overline{Y}_{X_2\cdot}$	$\frac{2+6+5}{3} = 4.(3)$	2.(6)	$\frac{2+1+6+3+5+4}{6} = 3.5$	
$\hat{\beta}_{2l}$	$4.(3) - 3.5 = 0.8(3)$	$-0.8(3)$		

$$\hat{Y}_1 = -2 + 0.8(3) + 3.5 = 2.(3) \qquad \hat{Y}_4 = 1 - 0.8(3) + 3.5 = 3.(6)$$
$$\hat{Y}_2 = -2 - 0.8(3) + 3.5 = 0.(6) \qquad \hat{Y}_5 = 1 + 0.8(3) + 3.5 = 5.(3)$$
$$\hat{Y}_3 = 1 + 0.8(3) + 3.5 = 5.(3) \qquad \hat{Y}_6 = 1 - 0.8(3) + 3.5 = 3.(6)$$

$n > 1$ хэрэглэгчээс авсан судалгааны мэдээлэлд үндэслэсэн загварын параметрийн үнэлгээ

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	$\frac{1+2+3+4+4+3}{6} = 2.83$	$2.83 - 3.5 = -0.66$
2	2,1	5,5	6,6	4.16	0.66
$\overline{Y}_{X_2\cdot}$	1.5	4.25	4.75	3.5	
$\hat{\beta}_{2l}$	-2	0.75	1.25		

Хэсэг 4

Энгийн шугаман регрессийн загварт хувиргах нь

Загварыг энгийн шугаман регресст 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.

Зайрмагтай 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 хүссэнээрээ 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

$$\begin{aligned} Y_1 &= \beta_{11} + \beta_{21} + \mu & Y_4 &= \beta_{12} + \beta_{22} + \mu \\ Y_2 &= \beta_{11} + \beta_{22} + \mu & Y_5 &= \beta_{13} + \beta_{21} + \mu \\ Y_3 &= \beta_{12} + \beta_{21} + \mu & Y_6 &= \beta_{13} + \beta_{22} + \mu \end{aligned}$$

болохыг анхаараад $\beta_1 = Y_6$ гэж эхэлбэл

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} Y_6 \\ Y_5 - Y_6 \\ Y_4 - Y_6 \\ Y_2 - Y_6 \end{pmatrix}$$

$$= \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ (\beta_{13} + \beta_{21} + \mu) - (\beta_{13} + \beta_{22} + \mu) \\ (\beta_{12} + \beta_{22} + \mu) - (\beta_{13} + \beta_{22} + \mu) \\ (\beta_{11} + \beta_{22} + \mu) - (\beta_{13} + \beta_{22} + \mu) \end{pmatrix} = \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ \beta_{21} - \beta_{22} \\ \beta_{12} - \beta_{13} \\ \beta_{11} - \beta_{13} \end{pmatrix}$$

болно.

X матрицыг олох байдал

$Y = X\beta$ буюу

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} \beta_{11} + \beta_{21} + \mu \\ \beta_{11} + \beta_{22} + \mu \\ \beta_{12} + \beta_{21} + \mu \\ \beta_{12} + \beta_{22} + \mu \\ \beta_{13} + \beta_{21} + \mu \\ \beta_{13} + \beta_{22} + \mu \end{pmatrix} = \begin{pmatrix} ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \\ ? & ? & ? & ? \end{pmatrix} \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ \beta_{21} - \beta_{22} \\ \beta_{12} - \beta_{13} \\ \beta_{11} - \beta_{13} \end{pmatrix}$$

байхыг анхаарна. Эндээс

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

гэж олдоно.

$Y = X\beta$ 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 $n = 1$ байх эхний тохиолдолд тус загварын параметрүүд

$$\hat\beta = \begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \\ \hat\beta_3 \\ \hat\beta_4 \end{pmatrix} = \begin{pmatrix} 3.(6) \\ 1.(6) \\ 0 \\ -3 \end{pmatrix}$$

гэж олдоно.

Шинжилгээний загварын β_{kl} параметрийг олох

$$\begin{cases} \hat\beta = \begin{pmatrix} 3.(6) \\ 1.(6) \\ 0 \\ -3 \end{pmatrix} = \begin{pmatrix} \beta_{13} + \beta_{22} + \mu \\ \beta_{21} - \beta_{22} \\ \beta_{12} - \beta_{13} \\ \beta_{11} - \beta_{13} \end{pmatrix} \\ \begin{cases} \beta_{11} + \beta_{12} + \beta_{13} = 0 \\ \beta_{21} + \beta_{22} = 0 \end{cases} \end{cases} \quad \begin{pmatrix} \hat\beta_{11} \\ \hat\beta_{12} \\ \hat\beta_{13} \\ \hat\beta_{21} \\ \hat\beta_{22} \\ \hat\mu \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0.8(3) \\ -0.8(3) \\ 3.5 \end{pmatrix}$$

```
conjoint::caPartUtilities(
  y = c(2,1,6,3,5,4),
  x = expand.grid(
    packing = c("biscuit", "plastic"),
    ingredients = c("curd", "chocolate", "fruit")),
  z = c("biscuit","plastic","curd","chocolate","fruit"))
```

```
intercept biscuit plastic curd chocolate fruit
      3.5   0.833  -0.833   -2         1     1
```


Хувилбаруудын эрэмбийн загвараар үнэлэгдсэн утга

$$\begin{pmatrix} \hat\beta_{11} \\ \hat\beta_{12} \\ \hat\beta_{13} \\ \hat\beta_{21} \\ \hat\beta_{22} \\ \hat\mu \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 1 \\ 0.8(3) \\ -0.8(3) \\ 3.5 \end{pmatrix}$$

$$\hat Y_1 = \hat\beta_{11} + \hat\beta_{21} + \hat\mu = -2 + 0.8(3) + 3.5 = 2.(3)$$
$$\hat Y_2 = \hat\beta_{11} + \hat\beta_{22} + \hat\mu = -2 - 0.8(3) + 3.5 = 0.(6)$$
$$\hat Y_3 = \hat\beta_{12} + \hat\beta_{21} + \hat\mu = 1 + 0.8(3) + 3.5 = 5.(3)$$
$$\hat Y_4 = \hat\beta_{12} + \hat\beta_{22} + \hat\mu = 1 - 0.8(3) + 3.5 = 3.(6)$$
$$\hat Y_5 = \hat\beta_{13} + \hat\beta_{21} + \hat\mu = 1 + 0.8(3) + 3.5 = 5.(3)$$
$$\hat Y_6 = \hat\beta_{13} + \hat\beta_{22} + \hat\mu = 1 - 0.8(3) + 3.5 = 3.(6)$$

Үүнийг бас $\hat Y = X\hat\beta$ байдлаар олох боломжтой.

$n > 1$ хэрэглэгчээс авсан судалгааны мэдээлэл дээр шинжилгээ хийх

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.

Лекц XV

Каноник корреляцын шинжилгээ

Каноник корреляцын шинжилгээ сэдвийн агуулга

1 Каноник корреляц

2 Шугаман эвлүүлгүүдийн корреляц

3 Каноник корреляцын тухай таамаглал

Хэсэг 1

Каноник корреляц

Каноник корреляцын шинжилгээ

Каноник корреляцыг хоёр санамсаргүй векторын холбоо хамаарлыг хэмжихэд ашигладаг.

№	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: 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

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

```
res <- CCA::cc(X, Y)
res$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

$$\max_{a,b} \rho(a^T X, b^T Y)$$

утгыг *каноник корреляцын коэффициент* гэнэ.

Хэсэг 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

$$\begin{aligned}\Sigma_{XX} &= \mathrm{cov}(X) && (q\times q)\\ \Sigma_{YY} &= \mathrm{cov}(Y) && (p\times p)\\ \Sigma_{XY} &= \Sigma_{YX}^T = \mathrm{cov}(X,Y) = E(X-\mu)(Y-\nu)^T && (q\times p)\end{aligned}$$

байна. Тэгвэл a^TX ба b^TY скаляр санамсаргүй хувьсагчдын корреляцын коэффициентыг

$$\begin{aligned}\rho(a^TX, b^TY) &= \frac{\mathrm{cov}(a^TX, b^TY)}{(\mathrm{cov}(a^TX))^{1/2}(\mathrm{cov}(b^TY))^{1/2}}\\ &= \frac{a^T\Sigma_{XY}b}{(a^T\Sigma_{XX}a)^{1/2}(b^T\Sigma_{YY}b)^{1/2}}\end{aligned}$$

байдлаар олж болно.

Каноник корреляцын бодлого

$\forall c \in \mathbb{R}^+$ тогтмолын хувьд $\rho(c\cdot\xi, \eta) = \rho(\xi, \eta)$ байдаг өөрөөр хэлбэл корреляцын коэффициент масштабаас хамаардаггүй тул

$$\begin{aligned}\mathrm{cov}(a^TX) &= a^T\Sigma_{XX}a = 1\\ \mathrm{cov}(b^TY) &= b^T\Sigma_{YY}b = 1\end{aligned}$$

буюу a^TX ба b^TY хувьсагчдын дисперсийг нэгтэй тэнцүү гэж тооцох боломжтой. Иймд каноник корреляцын коэффициент олохын тулд

$$\begin{aligned}\max_{a,b}\quad & a^T\Sigma_{XY}b\\ \text{s.t.}\quad & a^T\Sigma_{XX}a = 1,\\ & b^T\Sigma_{YY}b = 1\end{aligned}$$

заагалттай оптимизацын бодлого бодох шаардлага тулгарч байна.

$$\begin{aligned}\mathrm{cov}(a^TX) &= a^T\Sigma_{XX}a = 1\\ \mathrm{cov}(b^TY) &= b^T\Sigma_{YY}b = 1\end{aligned}$$

буюу a^TX болон b^TY хувьсагчдын дисперсийг нэгтэй тэнцүү гэсэн нөхцөлийг өөрөөр харъя. Санамсаргүй векторуудыг

$$\Sigma_{XX}^{-1/2}X$$
$$\Sigma_{YY}^{-1/2}Y$$

гэж хувиргасан гэж үзвэл шугаман эвлүүлгийн векторуудыг үүний урвуугаар $\Sigma_{XX}^{1/2}a$, $\Sigma_{YY}^{1/2}b$ гэж хувиргасан гэж тооцох ёстой болно. Тэгвэл хувиргалтаар үүсэх санамсаргүй векторуудын ковариац

$$\mathrm{cov}(\Sigma_{XX}^{-1/2}X, \Sigma_{YY}^{-1/2}Y) = \Sigma_{XX}^{-1/2}\,\mathrm{cov}(X,Y)\Sigma_{YY}^{-1/2} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$$

болно. Тэгэхээр бид бодлогынхоо 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 ковариацын 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

- $\Gamma = (\gamma_1, \ldots, \gamma_k)$, $\Delta = (\delta_1, \ldots, \delta_k)$, $\Lambda = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})$
- $k = \mathrm{rank}(K) \le \min\{q, p\}$
- $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_k$ 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 заагалтын нөхцөл хангах

$$\begin{aligned}a &= \Sigma_{XX}^{-1/2}\gamma_i\\ b &= \Sigma_{YY}^{-1/2}\delta_i\end{aligned}$$

орлуулга хийвэл зорилгын функцийн утга

$$\begin{aligned}a^T\Sigma_{XY}b &= a^T\Sigma_{XX}^{1/2}\Gamma\Lambda\Delta^T\Sigma_{YY}^{1/2}b\\ &= \gamma_1^T\Sigma_{XX}^{-1/2}\Sigma_{XX}^{1/2}\Gamma\Lambda\Delta^T\Sigma_{YY}^{1/2}\Sigma_{YY}^{-1/2}\delta_1 = \gamma_1^T\Gamma\Lambda\Delta^T\delta_1 = \lambda_i^{1/2}\end{aligned}$$

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 $i \ne j$ үед

$$\begin{aligned}\gamma_i^T\gamma_j &= a_i^T\Sigma_{XX}^{1/2}\Sigma_{XX}^{1/2}a_j = a_i^T\Sigma_{XX}a_j = 0\\ \delta_i^T\delta_j &= b_i\Sigma_{YY}^{1/2}\Sigma_{YY}^{1/2}b_j = b_i\Sigma_{YY}b_j = 0\end{aligned}$$

байна. Мөн бодлогыг анх томьёолохдоо тавьсан заагалт энд хамаарна.

Бодлогын шийд

Олсон шийдээ теорем байдлаар томьёолж бичье.

Теорем

Өгсөн r $(1 \le r \le k)$ 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

$$\begin{aligned}a_i &= \Sigma_{XX}^{-1/2}\gamma_i\\ b_i &= \Sigma_{YY}^{-1/2}\delta_i\end{aligned}$$

- Шугаман эвлүүлгээр үүсэх хувьсагч

$$\begin{aligned}\eta_i &= a_i^TX = \gamma_i^T\Sigma_{XX}^{-1/2}X\\ \varphi_i &= b_i^TY = \delta_i^T\Sigma_{YY}^{-1/2}Y\end{aligned}$$

  Эдгээрийг *каноник хувьсагч* гэнэ. Хэрэв $EX = 0$ ба $EY = 0$ нөхцөл тавьсан гэвэл эндээс олдох каноник хувьсагчийн утга CCA багцын `cc()` функцийнхтэй адил болно.
- Каноник корреляцын коэффициент

$$\rho(\eta_i, \varphi_i) = \rho(a_i^TX, b_i^TY) = \lambda_i^{1/2}$$

Каноник корреляцын чанар

Чанар

1. Каноник хувьсагчдын ковариац дараах байдалтай байна.

$$\mathrm{cov}(\eta, \varphi) = \begin{pmatrix}I_k & \Lambda\\ \Lambda & I_k\end{pmatrix}$$

   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. $$\mathrm{cov}(\eta_i,\eta_j)=\mathrm{cov}(a_i^TX,a_j^TX)=a_i^T\,\mathrm{cov}(X,X)a_j=a_i^T\Sigma_{XX}a_j$$
$$=\gamma_i^T\Sigma_{XX}^{-1/2}\Sigma_{XX}\Sigma_{XX}^{-1/2}\gamma_j=\gamma_i^T\gamma_j=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$$

$\mathrm{cov}(\varphi_i,\varphi_j)$ ковариацад харгалзах тооцоо үүнтэй төстэй.

$$\mathrm{cov}(\eta_i,\varphi_j)=\mathrm{cov}(a_i^TX,b_j^TY)=a_i^T\,\mathrm{cov}(X,Y)b_j=a_i^T\Sigma_{XY}b_j$$
$$=\gamma_i^T\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta_j=\gamma_i^TK\delta_j=\gamma_i^T\Gamma\Lambda\Delta^T\delta_j=\lambda_{ij}$$

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. □

Хэсэг 3

Каноник корреляцын тухай таамаглал

Каноник корреляцын тухай таамаглалууд

1. Каноник корреляцын бүх коэффициент тэгтэй тэнцүү буюу X болон Y санамсаргүй векторууд хамааралгүй
$$-\{n-(p+q+3)/2\}\ln\prod_{i=1}^{k}(1-\lambda_i)\sim\chi^2_{pq}$$
2. Каноник корреляцын коэффициентуудын зөвхөн эхний s ширхэг нь тэгээс ялгаатай
$$-\{n-(p+q+3)/2\}\ln\prod_{i=s+1}^{k}(1-\lambda_i)\sim\chi^2_{(p-s)(q-s)}$$

Жишээ

Өмнө авсан жишээний хувьд дээрх хоёр таамаглалыг 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 тус таамаглалуудыг дараах байдлаар шалгана.

1. $H_0:\lambda_1^{1/2}=\lambda_2^{1/2}=0$
$$-\{12-(2+3+3)/2\}\ln\{(1-0.84^2)(1-0.62^2)\}=13.95$$
p-утга $=1-\chi^2_6(13.95)=0.03<\alpha=0.05$ тул тэг таамаглал няцаагдана.
2. $H_0:s=1$ буюу $\lambda_1^{1/2}\neq0$, $\lambda_2^{1/2}=0$
$$-\{12-(2+3+3)/2\}\ln(1-0.62^2)=3.92$$
p-утга $=1-\chi^2_2(3.92)=0.14>\alpha=0.05$ тул тэг таамаглалыг няцаах үндэслэлгүй.

Ном зүй

Г.Махгал, Ш.Мөнгөнсүх
Олон хэмжээст өгөгдлийн статистик шинжилгээ
2017.