다변량통계학-2025년 2학기

서울시립대학교 통계학과 이용희

2025-10-29

Table of contents

List of Figures

List of Tables

Preface

이 책은 2025년 다변량통계학에 대한 온라인 교재입니다.

> **표기법**
>
> 이 책에서 사용된 기호, 표기법, 프로그램의 규칙과 쓰임은 다음과 같습니다.
>
> - 스칼라(scalar)와 일변량 확률변수는 일반적으로 보통 글씨체의 소문자로 표기한다. 특별한 이유가 있는 경우 대문자로 표시할 것이다.
> - 벡터, 행렬, 다변량 확률벡터는 **굵은 글씨체**로 표기한다.
> - 통계 프로그램은 `R`을 이용하였다. 각 예제에 사용된 `R` 프로그램은 코드 상자를 열면 나타난다.

1. 확률벡터와 다변량 정규분포

```
library(Hotelling)
library(tidyverse)
library(here)
library(knitr)
library(purrr)
library(rmarkdown)
library(kableExtra)
library(flextable)
```

다변량 자료(multivariate data)는 두 개 이상의 변수를 측정한 자료를 말합니다. 예를 들어, 학생들의 키와 몸무게, 시험 점수와 공부 시간, 나이와 소득 등이 다변량 자료에 해당합니다. 다변량 자료는 변수들 간의 관계를 분석하고 이해하는 데 중요한 역할을 합니다. 다변량 자료를 효과적으로 표현하고 분석하기 위해 다양한 그래프와 통계 기법이 사용됩니다. 이 장에서는 다변량 자료의 표현 방법과 분포를 이해하는 데 필요한 기본 개념과 도구들을 소개합니다.

1.1. 예제- 국민체력100

국민체력100은 국민의 체력증진과 건강증진을 위해 개발된 종합적인 체력측정 프로그램이다. 이 프로그램은 다양한 연령대와 성별에 맞춘 체력측정 항목을 포함하고 있으며, 이를 통해 개인의 체력 상태를 평가하고 개선할 수 있는 기회를 제공한다.

이번 장에서는 청소년(13-18세) 남여 3000명에 대하여 2024년에 국민체력100 사업에서 측정한 자료를 예제로 사용하여 다변량 자료를 표현하는 방법들과 분포를 배울것이다.

먼저 측정항목에 대한 설명에 대한 자료를 보자.

```
load(here("data", "physical100.RData"))
ls()
```

```
[1] "physical100_df"      "physical100_df_info"
```

먼저 데이터프레임 `selected_var_df` 에는 측정한 항목의 영문 변수이름(`varname_eng`), 종목의 설명(`varname_kor`), 측정분야(`category_kor`) 그리고 측정단위(`unit`) 가 다음과 같이 저장되어 있다.

varname_eng	varname_kor	category_kor	unit
height	신장	신체구성	cm
weight	체중	신체구성	kg
body_fat_pct	체지방율	신체구성	
grip_left	악력_좌	근력	kg
grip_right	악력_우	근력	kg
sit_forward	앉아윗몸앞으로굽히기	유연성	cm
illinois	일리노이	민첩성	초
hang_time	청소년체공시간	순발력	초
twall_time	TWALL_시간	협응력	초
twall_errors	TWALL_실수	협응력	회
twall_score	TWALL_결과값	협응력	초
bmi	BMI	신체구성	
rel_grip	상대악력	근력	%
abs_grip	절대악력	근력	kg

다음으로 청소년 3000명의 측정 자료의 일부는 다음과 같다.

sex	age	height	weight	body_fat_pct	grip_left	sit_forward
남성	15	166.5	68.0	26.3	31.9	22.1
여성	13	166.4	45.5	22.0	20.0	10.2
남성	13	163.2	44.7	11.7	22.0	-2.0
여성	14	156.9	44.7	26.9	17.3	-5.0
남성	17	175.7	78.1	16.7	52.2	18.5
여성	16	167.2	74.5	37.1	25.9	12.0
여성	16	162.0	57.3	37.1	21.6	-4.5
여성	17	169.1	75.0	39.8	21.6	0.1

sex	age	height	weight	body_fat_pct	grip_left	sit_forward
여성	15	160.9	56.8	33.1	25.1	-8.0
남성	13	162.8	57.6	18.0	39.6	28.0

1.2. 확률벡터와 기본 성질

1.2.1. 일변량 확률변수

일변량 확률변수(random variable) X가 확률밀도함수 $f(x)$를 가지는 분포를 따를때 기대값과 분산은 다음과 같이 정의된다.

$$E(X) = \int x f(x) dx = \mu$$
$$V(X) = E[X - E(X)]^2 = \int (x-\mu)^2 f(x) dx = \sigma^2$$

새로운 확률변수 Y가 확률변수 X의 다음과 같은 선형변환으로 표시된다면 (a와 b는 실수)

$$Y = aX + b$$

일변량 확률변수 X의 기대값(평균)과 분산은 다음과 같이 계산된다.

$$\begin{aligned} E(Y) &= E(aX+b) \\ &= \int (ax+b) f(x) dx \\ &= a \int x f(x) dx + b \\ &= aE(X) + b \\ &= a\mu + b \end{aligned}$$

$$\begin{aligned} V(Y) &= Var(aX+b) \\ &= E[aX + b - E(aX+b)]^2 \\ &= E[a(X-\mu)]^2 \\ &= a^2 E(X-\mu)^2 \\ &= a^2 \sigma^2 \end{aligned}$$

1.2.2. 다변량 확률벡터

이제 하나의 학률변수가 아는 2개 이상의 확률변수들을 모아놓은 확률벡터(random vector)를 생각해 보자. 다음과 같이 벡터로 표현된 확률벡터 $\boldsymbol{X}$가 p 차원의 다변량 분포(multivariate distribution)를 따른다고 하고 결합확률밀도함수 $f(\boldsymbol{x}) = f(x_1, x_2, \dots, x_p)$를 를 가진다고 하자.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ .. \\ X_p \end{bmatrix}$$

다변량 확률벡터의 평균 벡터(mean vector)는 다음과 같이 주어진다. 확률벡터의 평균 벡터는 구성하는 각 확률변수의 평균으로 주어진다.

$$\boldsymbol{E}(\boldsymbol{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \\ .. \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ .. \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

다음으로 공분산(covariance)과 상관계수(correlation coefficient)에 대해서 알아보자. 우리는 여러 개의 확률변수의 관계를 분석하는 분석을 하려고 하는데, 이 경우 가장 많이 사용되는 통계량이 두 개의 변수들의 선형적 관계를 나타내는 상관계수이다. 두 확률변수 X_k 와 X_l 의 상관계수 ρ_{jk} 는 다음과 같이 정의된다.

$$\rho_{jk} = \frac{Cov(X_j, X_k)}{\sqrt{V(X_j)V(X_k)}} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}, \quad j, k = 1, 2, \dots, p$$

위의 상관계수의 공식에서 $\sigma_{jj} \equiv \sigma_j^2$ 와 $\sigma_{kk} = \sigma_k^2$ 는 각각 확률변수 X_i 와 X_j 의 분산이며, 공분산은 다음과 같이 정의된다.

$$\begin{aligned} Cov(X_j, X_k) &= E[(X_j - E(X_j))(X_k - E(X_k))] \\ &= E(X_j X_k) - E(X_j)E(X_k) \end{aligned}$$

위의 식을 보면 각각의 확률 변수가 평균에서 차이가 나는 두 개의 편차, 즉 $X_j - E(X_j)$, $X_k - E(X_k)$ 의 곱에 대한 기대값으로 두 확률 변수가 평균에서 얼마나 같은 방향 또는 반대 방향으로 함께 움직이는 경향이 있는지 그 정도를 수치화한 값이다. 두 확률변수의 공분산의 값이 양의 값으로 커지면 두 확률 변수의 변화가 같은 방향으로 나타난다는 의미이며, 반대로 음의 값으로 커지면 두 확률 변수의 변화가 반대 방향으로 나타난다는 의미이다.

참고로 공분산은 단위가 확률 변수의 단위에 영향을 받기 때문에 크기 자체만으로 비교가 직관적이지 않다는 단점이 있다. 반면에 상관 계수는 공분산을 각 확률 변수의 표분편차로 나누어 얻은 값이므로 단위에 영향을 받지 않아서 상대적인 비교가 가능하다.

상관계수는 -1 과 1 사이의 값을 가지며 1에 가까울수록 두 개의 변수가 같은 방향으로 움직이는 확률적 경향이 강해지며 반대로 -1 에 가까워질수록 반대의 방향을 움직이는 경향이 강해진다.

여기서 중요한 점은 상관계수(또는 공분산)은 두 확률 변수의 선형적 관계(linear relationship)을 나타내는 통계량으로 비선형적 관계를 파악하는데는 한계가 있을 수 있다.

이제 확률 벡터의 모든 변수에 대한 분산과 공분산을 다음과 같은 공분산 행렬로 나타낼 수 있다.

$$\begin{aligned} V(\boldsymbol{X}) = Cov(\boldsymbol{X}) &= E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t \\ &= E(\boldsymbol{X}\boldsymbol{X}^t) - \boldsymbol{\mu}\boldsymbol{\mu}^t \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ & \cdots & \cdots & \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} \\ &= \boldsymbol{\Sigma} \end{aligned}$$

여기서 $\sigma_{jj} = V(X_j)$, $\sigma_{jk} = Cov(X_j, X_k) = Cov(X_k, X_j)$ 이다. 따라서 공분산 행렬 $\boldsymbol{\Sigma}$는 대칭행렬(symmetric matrix)이다.

더 나아가 확률 벡터의 모든 변수에 대한 상관계수을 다음과 같은 상관계수 행렬(correlation matrix) $\boldsymbol{R}$ 로 나타낼 수 있다.

$$\begin{aligned} cor(\boldsymbol{X}) &= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ & \cdots & \cdots & \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \\ &= \boldsymbol{R} \end{aligned}$$

위의 상관계수 행렬에서 대각원소는 모두 1 임이 유의하자.

새로운 확률벡터 $\boldsymbol{Y}$가 확률벡터 $\boldsymbol{X}$ 의 선형변환라고 하자.

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$$

단 여기서 $\boldsymbol{A}$ 는 $p \times p$ 실수 행렬이고 $\boldsymbol{b}$ 는 $p \times 1$ 실수 벡터이다.

확률벡터 $\boldsymbol{Y}$ 의 기대값(평균벡터)과 공분산은 다음과 같이 계산된다.

$$
\begin{aligned}
E(\boldsymbol{Y}) &= E(\boldsymbol{AX}+\boldsymbol{b}) \\
&= \boldsymbol{A}E(\boldsymbol{X})+\boldsymbol{b} \\
&= \boldsymbol{A\mu}+\boldsymbol{b} \\
V(\boldsymbol{Y}) &= Var(\boldsymbol{AX}+\boldsymbol{b}) \\
&= E[\boldsymbol{AX}+\boldsymbol{b}-E(\boldsymbol{AX}+\boldsymbol{b})][\boldsymbol{AX}+\boldsymbol{b}-E(\boldsymbol{AX}+\boldsymbol{b})]^t \\
&= E[\boldsymbol{AX}-\boldsymbol{A\mu}][\boldsymbol{AX}-\boldsymbol{A\mu}]^t \\
&= E[\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})][\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})]^t \\
&= E[\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t\boldsymbol{A}^t] \\
&= \boldsymbol{A}E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t]\boldsymbol{A}^t \\
&= \boldsymbol{A\Sigma A}^t
\end{aligned}
$$

1.2.3. 표본 통계량

이제 확률 표본(sample)을 이용하여 평균벡터, 공분산, 상관계수를 추정하는 간단한 방법에 대해서 알아보자.

확률 벡터 $\boldsymbol{X}$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 다변량 분포 F 를 따른다고 가정하자. 만약 확률 표본 $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ 이 독립적으로 다변량 분포 F 에서 임의로 추출되었다면

$$
\boldsymbol{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ X_{i3} \\ .. \\ X_{ip} \end{bmatrix} \quad i = 1, 2, \dots, n
$$

다음과 같이 표본 통계량을 이용하여 평균벡터, 공분산, 상관계수를 추정할 수 있다.

먼저 다음과 같은 표본평균 벡터 $\bar{\boldsymbol{X}}$ 는 평균벡터 $\boldsymbol{\mu}$ 의 불편추정량(unbiased estimator)이다.

$$
\bar{\boldsymbol{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ .. \\ \bar{X}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n X_{i1}/n \\ \sum_{i=1}^n X_{i2}/n \\ \sum_{i=1}^n X_{i3}/n \\ .. \\ \sum_{i=1}^n X_{ip}/n \end{bmatrix} = \hat{\boldsymbol{\mu}}
$$

여기서 X_{ij} 는 i 번째 표본벡터 $\boldsymbol{X}_i = (X_{i1} X_{i2} \dots X_{ip})^t$ 의 j 번째 확률변수이다.

또한 아래에 주어진 표본 공분산 행렬 $\boldsymbol{S}$ 은 공분산 행렬 $\boldsymbol{\Sigma}$ 의 추정량이다.

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{12} & s_{22} & \dots & s_{2p} \\ & \dots & \dots & \\ s_{1p} & s_{2p} & \dots & s_{pp} \end{bmatrix} = \hat{\boldsymbol{\Sigma}}$$

위에서 $s_{jj} \equiv s_j^2$ 는 확률변수 X_j 의 표본 분산이며 s_{jk} 는 X_j 와 X_k 의 표본 공분산이며 다음과 같이 계산된다.

$$s_{jj} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2, \quad s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k), \quad j,k = 1,2,\dots,p$$

마지막으로 아래에 주어진 표본 상관계수 행렬 $\boldsymbol{R}$ 은 상관계수 행렬 $\boldsymbol{R}$ 의 추정량이다.

$$\hat{\boldsymbol{R}} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ & \dots & \dots & \\ r_{1p} & r_{2p} & \dots & 1 \end{bmatrix}$$

여기서 r_{jk} 는 확률변수 X_j 와 X_k 의 표본 상관계수이며 다음과 같이 계산된다.

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}, \quad j,k = 1,2,\dots,p$$

1.2.4. 예제-국민체력100

이제 위에서 살펴본 국민체력100 자료에서 청소년 남자 자료를 이용하여 평균벡터, 공분산 행렬, 상관계수 행렬의 표본 통계량을 계산해 보자.

먼저 표본 평균 벡터를 계산해 보자. 주어진 변수가 많으니 키(`height`), 몸무게(`weight`), 체지방률(`body_fat_pct`), 악력(`grip_left`), 앉아윗몸앞으로굽히기(`sit_forward`), 청소년체공시간(`hang_time`) 6개 변수만 선택하여 계산해 보자.

```
# 국민체력100 자료에서 남자만 선택하여 데이터프레임 df 생성
df <- physical100_df %>%
  filter(sex == "남성") %>%
  select(height, weight, body_fat_pct, grip_left, sit_forward,hang_time)

#  패키지 dplyr의 summarise()와 across() 함수를 사용하여 각 열의 평균 계산
sample_mean_vector <- df %>%
  summarise(across(everything(), \(x) mean(x, na.rm = TRUE))) %>%
  unlist()

sample_mean_vector
```

```
      height       weight body_fat_pct    grip_left  sit_forward    hang_time
 172.0451492   69.7115470   20.9993923   36.0728729    7.8976630    0.5593964
```

다음으로 표본 공분산 행렬을 계산해 보자.

```
cor(df)
```

```
                  height      weight body_fat_pct   grip_left   sit_forward
height        1.0000000000  0.50994161  -0.03019400  0.45225668  0.0008938941
weight        0.5099416134  1.00000000   0.69156075  0.46152254  0.0111910927
body_fat_pct -0.0301940037  0.69156075   1.00000000 -0.01208826 -0.1434596849
grip_left     0.4522566754  0.46152254  -0.01208826  1.00000000  0.2605654264
sit_forward   0.0008938941  0.01119109  -0.14345968  0.26056543  1.0000000000
hang_time     0.1884978140 -0.14265212  -0.48446817  0.34559521  0.2889235491
              hang_time
height        0.1884978
weight       -0.1426521
body_fat_pct -0.4844682
grip_left     0.3455952
sit_forward   0.2889235
hang_time     1.0000000
```

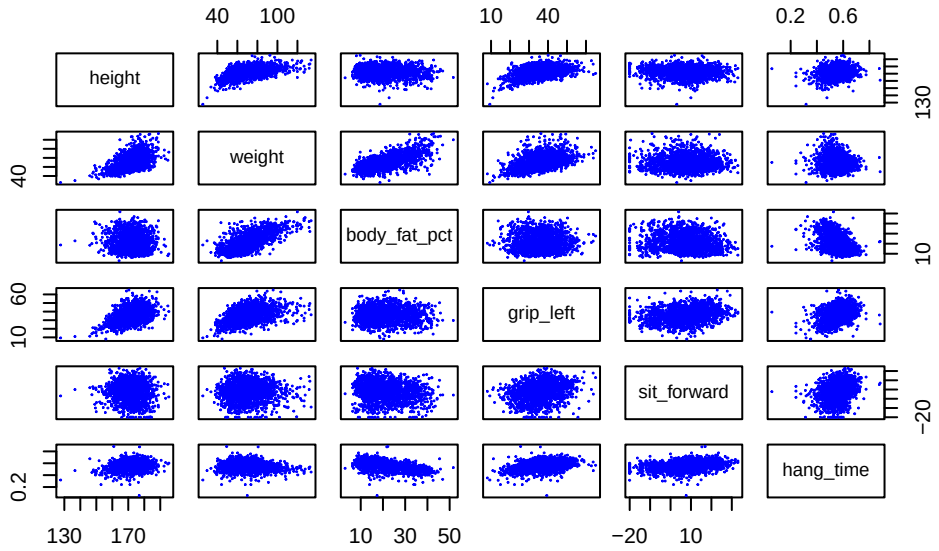
마지막으로 표본 상관계수 행렬을 계산해 보자.

```
cor(df)
```

```
                  height      weight body_fat_pct   grip_left   sit_forward
height        1.0000000000  0.50994161  -0.03019400  0.45225668  0.0008938941
weight        0.5099416134  1.00000000   0.69156075  0.46152254  0.0111910927
body_fat_pct -0.0301940037  0.69156075   1.00000000 -0.01208826 -0.1434596849
grip_left     0.4522566754  0.46152254  -0.01208826  1.00000000  0.2605654264
sit_forward   0.0008938941  0.01119109  -0.14345968  0.26056543  1.0000000000
hang_time     0.1884978140 -0.14265212  -0.48446817  0.34559521  0.2889235491
              hang_time
height        0.1884978
weight       -0.1426521
body_fat_pct -0.4844682
grip_left     0.3455952
sit_forward   0.2889235
hang_time     1.0000000
```

표본 상관계수 행렬을 보면 다양한 상관관계가 나타나는데 이러한 관계를 더 자세하게 보기위하여 산점도 행렬(scatterplot matrix)로 시각화 하면 더 유용한 정보를 얻을 수 있다.

```
pairs(df, pch=19, col='blue', cex=0.1)
```



1.3. 다변량 정규분포

일변량 확률변수 X가 평균이 μ 이고 분산이 σ^2인 정규분포를 따른다면 다음과 같이 나타내고

$$X \sim N(\mu, \sigma^2)$$

확률밀도함수 $f(x)$ 는 다음과 같이 주어진다.

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

1.3.1. 확률 밀도 함수

p-차원 확률벡터 $\boldsymbol{X}$가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$인 다변량 정규분포를 따른다면 다음과 같이 나타내고

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

확률밀도함수 $f(\boldsymbol{x})$ 는 다음과 같이 주어진다.

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

예를 들어 2-차원 확률벡터 $\boldsymbol{X} = (X_1, X_2)^t$가 평균이 $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ 이고 공분산 $\boldsymbol{\Sigma}$가 다음과 같이 주어진

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

이변량 정규분포를 따른다면 확률밀도함수 $f(\boldsymbol{x})$에서 exp함수의 인자는 다음과 같이 주어진다.

$$(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^t = \\ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}}\right)+\left(\frac{(x_2-\mu_2)^2}{\sigma_{22}}\right)-2\rho\left(\frac{(x_1-\mu_1)}{\sqrt{\sigma_{11}}}\right)\left(\frac{(x_2-\mu_2)}{\sqrt{\sigma_{22}}}\right)\right]$$

그리고 $p=2$ 인 경우 확률밀도함수의 상수부분은 다음과 같이 주어진다.

$$(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}}$$

여기서 $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$

> **다변량 정규분포에서 독립과 공분산**
>
> 다변량 정규분포에서 공분산이 0인 두 확률 변수는 독립이다.
>
> $$\sigma_{ij} = 0 \leftrightarrow X_i \text{ and } X_j \text{ are independent}$$
>
> 참고로 정규분포가 아닌 다른 분포의 경우 공분산이 0인 두 확률 변수는 독립이 아닐 수 있다.

1.3.2. 예제-국민체력100

이제 위에서 살펴본 국민체력100 자료에서 청소년 남자의 키(height)와 몸무게(weight) 가 이변량 정규분포를 따른다고 가정하고 확률밀도 함수를 그려보자.

```
# 필요한 패키지 로드
library(mvtnorm)
library(plotly)

# 국민체력100 자료에서 남자만 선택하여 데이터프레임 df 생성, 키와 몸무게 변수만 선택와
df <- physical100_df %>%
  filter(sex == "남성") %>%
  select(height, weight)

#  패키지 dplyr의 summarise()와 across() 함수를 사용하여 각 열의 평균 계산
sample_mean_vector <- df %>%
  summarise(across(everything(), \(x) mean(x, na.rm = TRUE))) %>%
```

```
  unlist()
sample_mean_vector
```

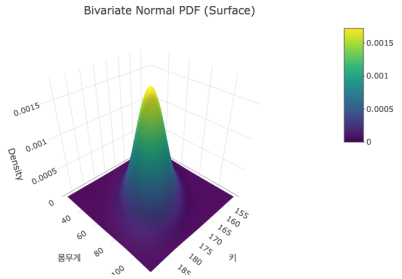
```
   height    weight
172.04515  69.71155
```

```
# 표본 공분산 행렬 계산
sample_cov_matrix <- cov(df, use = "complete.obs")
sample_cov_matrix
```

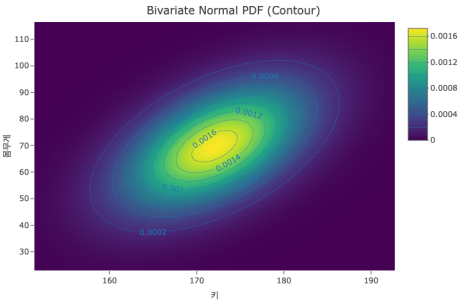
```
          height    weight
height 47.51166  54.80339
weight 54.80339 243.09372
```

```
# 이변량 정규분포의 확률밀도함수 계산
# 키와 몸무게의 평균에서 표분편차 3배의 범위의 값을 100개로 나누어 x,y 축 생성
x1_seq <- seq(sample_mean_vector[1]-3* sqrt(sample_cov_matrix[1,1]), sample_mean_vector[1]+3*
x2_seq <- seq(sample_mean_vector[2]-3* sqrt(sample_cov_matrix[2,2]), sample_mean_vector[2]+3*
grid <- expand.grid(height = x1_seq, weight = x2_seq)

# 확률밀도함수 계산 (z축의 값)
grid$z <- dmvnorm(grid, mean = sample_mean_vector, sigma = sample_cov_matrix)

# z를 행렬로 변환 (surface plot용)
z_matrix <- matrix(grid$z, nrow = length(x1_seq), ncol = length(x2_seq))
```

다음은 위에서 얻어진 확률밀도 함수를 3차원 surface plot으로 나타낸 것이다.



아래 그림은 표본으로 부터 얻어진 확률밀도 함수를 2차원 등고선(contour)으로 나타낸 그림이다.



1.3.3. 조건부 분포

다변량 정규분포 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 따르는 확률벡터 $\boldsymbol{X}$를 다음과 같이 두 부분으로 나누면

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \quad \boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{12} \\ \vdots \\ \boldsymbol{X}_{1p} \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} \boldsymbol{X}_{21} \\ \boldsymbol{X}_{22} \\ \vdots \\ \boldsymbol{X}_{2q} \end{bmatrix}$$

각각 다변량 정규분포를 따르고 다음과 같이 나타낼 수 있다.

$$\begin{bmatrix} E(\boldsymbol{X}_1) \\ E(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \begin{bmatrix} V(\boldsymbol{X}_1) & Cov(\boldsymbol{X}_1, X_2) \\ Cov(\boldsymbol{X}_2 X_1) & V(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^t & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N_{p+q} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \Sigma_{12} \\ \boldsymbol{\Sigma}_{12}^t & \Sigma_{22} \end{bmatrix} \right)$$

확률벡터 $\boldsymbol{X}_2 = \boldsymbol{x}_2$가 주어진 경우 $\boldsymbol{X}_1$의 조건부 분포는 p-차원 다변량 정규분포를 따르고 평균과 공분산은 다음과 같다.

$$E(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{x}_2), \quad V(\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t$$

만약 $X_2 = x_2$가 주어졌을 때 X_1의 조건부 분포는 정규분포이고 평균과 분산은 다음과 같이 주어진다.

$$E(X_1 | X_2 = x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(\mu_2 - x_2) = \mu_1 + \rho\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(\mu_2 - x_2)$$

$$V(X_1 | X_2 = x_2) = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma_{11}(1 - \rho^2)$$

2. 다변량 자료의 시각화

```
library(gapminder)
library(GGally)
library(pheatmap)

library(tidyverse)
library(here)
library(knitr)
library(rmarkdown)
library(kableExtra)
library(flextable)

#아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
#font_add_google("Nanum Pen Script", "gl")
font_add_google(name = "Noto Sans KR", family = "noto")
showtext_auto()
```

일반적으로 많이 사용하는 회귀모형(regression model) 은 반응변수와 이에 영향을 주는 설명변수들의 관계를 분석하는 모형이다. 따라서 주로 회귀모형에서 시각화 방법은 반응변수와 설명변수의 관계를 나타내는 그림이며 주로 2개의 변수의 관계를 파악하는 산점도 그림 (scatter plot) 이 많이 사용된다.

다변량 분석은 여러 개의 확률 변수들의 관계를 분석하는 통계 기법이다. 회귀모형과 다른 점은 주로 예측에 관심이 있는 반응변수가 없는 것이며 따라서 분석의 특성상 3개 이상의 변수의 관계를 한 그림에 시각화하는 방밥이 필요하다.

이 장에서는 다변량 분석에서 여러 개의 변수들의 관계를 동시에 시각화 방법들을 알아보려고 한다.

2.1. 산점도 그림

앞에서 언급한 것과 같이 3개의 변수를 산점도에 나타내려면 x 축, y 축 과 더불어서 점의 특성(색깔, 모양 등)을 이용할 수 있다.

R 패키지 gapminder 에 포함된 gapminder 데이터는 전 세계 여러 국가의 경제, 보건 지표를 연도별로 기록한 공개 자료이다. 원본은 [Gapminder 재단]{https://www.gapminder.org/} 에서 제공한다. gapminder 데이터

는 1952년부터 2007년까지 5년 단위로 측정한 각 국가별 경제 수준과 건강 상태를 시계열로 비교 가능하게 정리한 자료이다.

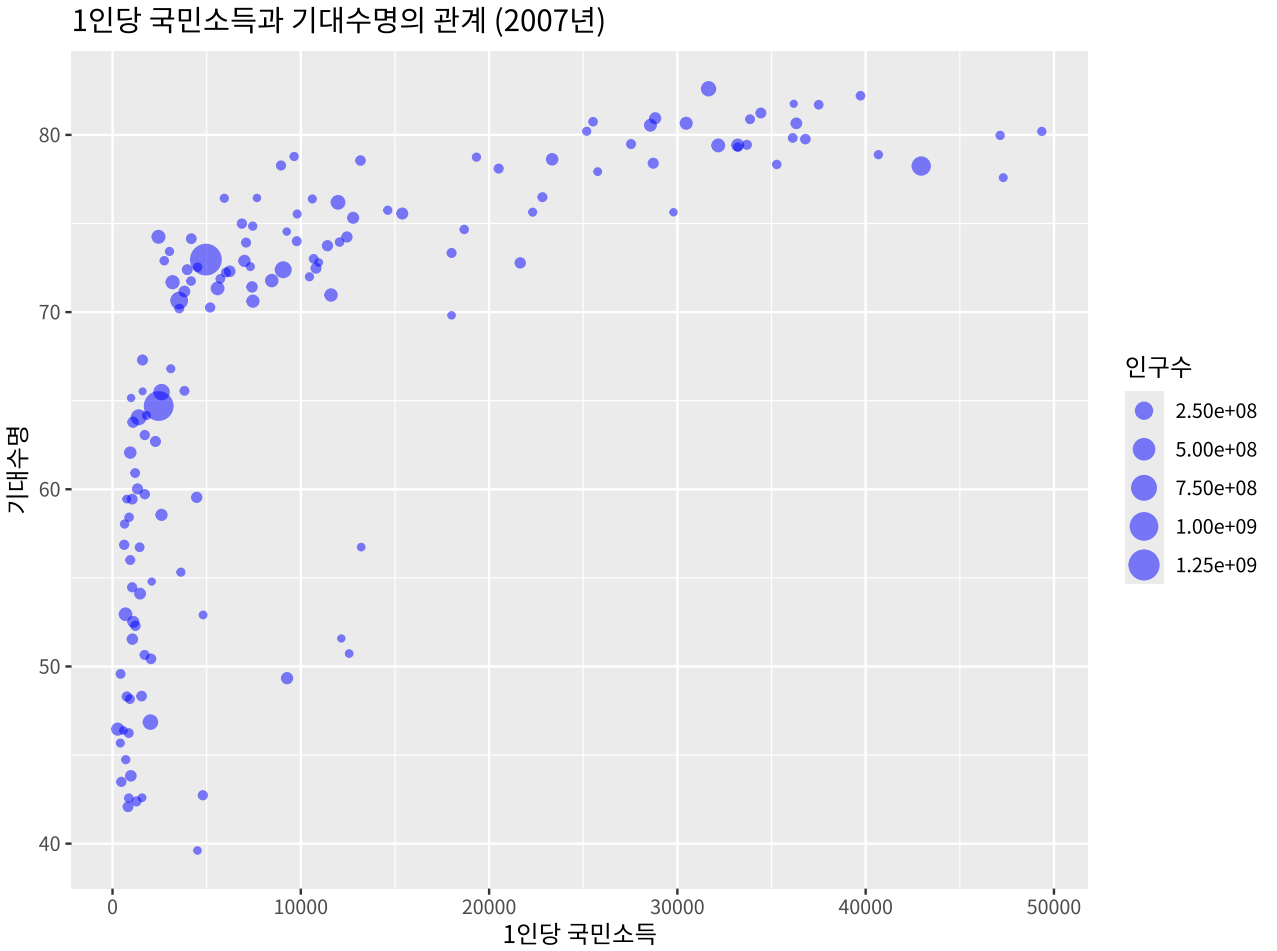
변수명	설명
`country`	국가 이름
`continent`	대륙 이름
`year`	연도
`lifeExp`	기대수명 (average life expectancy)
`pop`	인구 (population)
`gdpPercap`	1인당 국민소득 (gross domestic product per capita)

```
head(gapminder::gapminder)
```

```
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333      779.
2 Afghanistan Asia       1957    30.3  9240934      821.
3 Afghanistan Asia       1962    32.0 10267083      853.
4 Afghanistan Asia       1967    34.0 11537966      836.
5 Afghanistan Asia       1972    36.1 13079460      740.
6 Afghanistan Asia       1977    38.4 14880372      786.
```

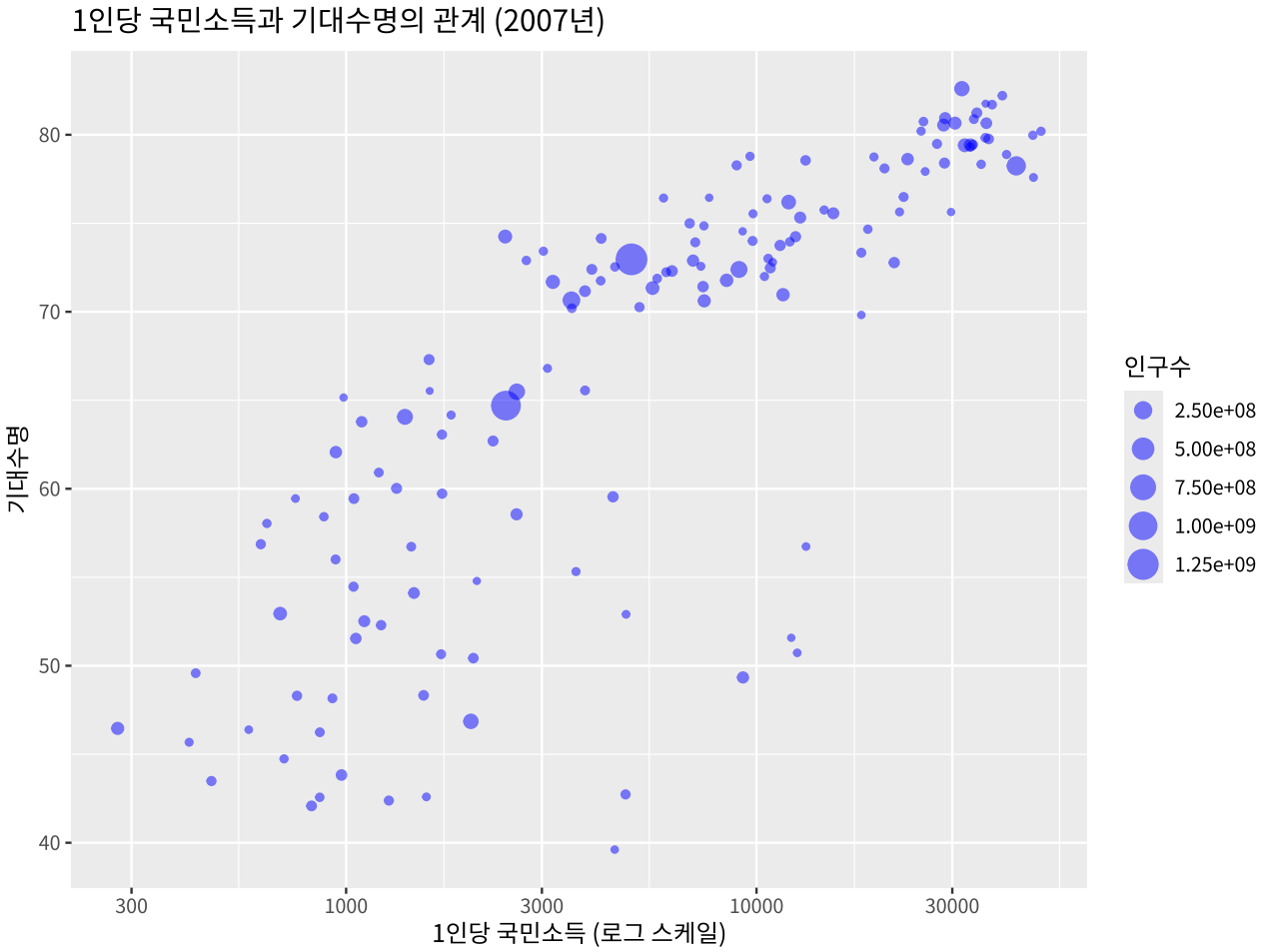
이제 `gapminder` 데이터에서 2007년 자료를 이용하여 1인당 국민소득(`gdpPercap`)과 기대수명(`lifeExp`)의 관계를 산점도로 나타내고, 대륙(`continent`)에 따라 점의 크기로 다르게 나타내는 그림을 그려보자. 이렇게 점의 크기를 변수의 값에 따라 변하는 산점도를 **버블 차트(bubble chart)** 라고 한다.

```
gapminder::gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, size=pop)) +
  geom_point(alpha = 0.5, color="blue") +
  labs(title = "1인당 국민소득과 기대수명의 관계 (2007년)",
       x = "1인당 국민소득",
       y = "기대수명",
       size = "인구수") +
  theme(text = element_text(family = "noto")) # 한글 폰트 설정(lib(showtext) 패키지 필요)
```



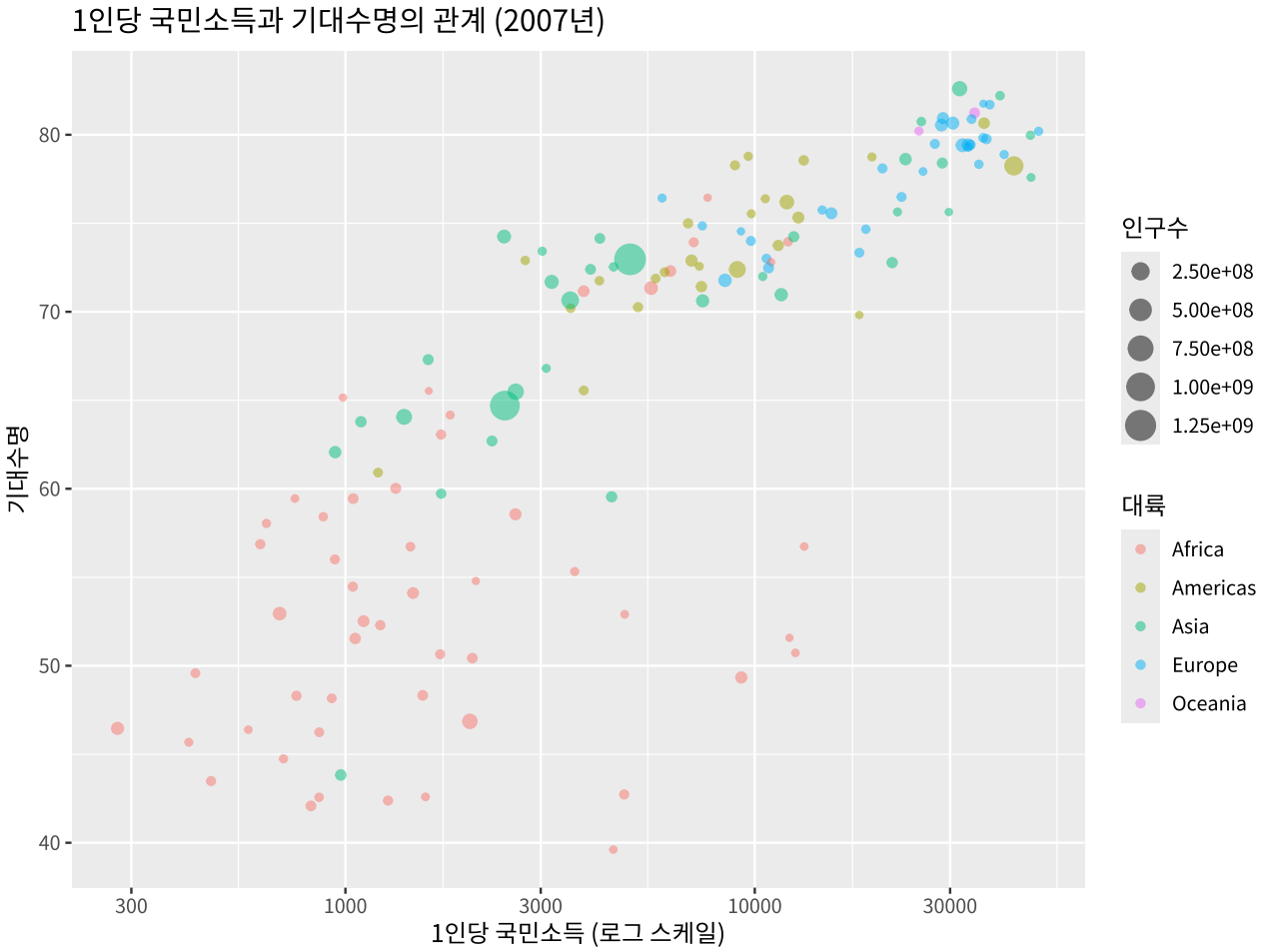
위의 그림을 보면 1인당 국민소득과 기대수명의 관계가 선형적이지 않음을 알 수 있다. 따라서 x 축을 로그 스케일로 변환하여 다시 그려보자.

```
gapminder::gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, size=pop)) +
  geom_point(alpha = 0.5, color="blue") + # alpha는 점의 투명도
  scale_x_log10() + # x축을 로그 스케일로 변환
  labs(title = "1인당 국민소득과 기대수명의 관계 (2007년)",
       x = "1인당 국민소득 (로그 스케일)",
       y = "기대수명",
       size = "인구수") +
  theme(text = element_text(family = "noto"))
```



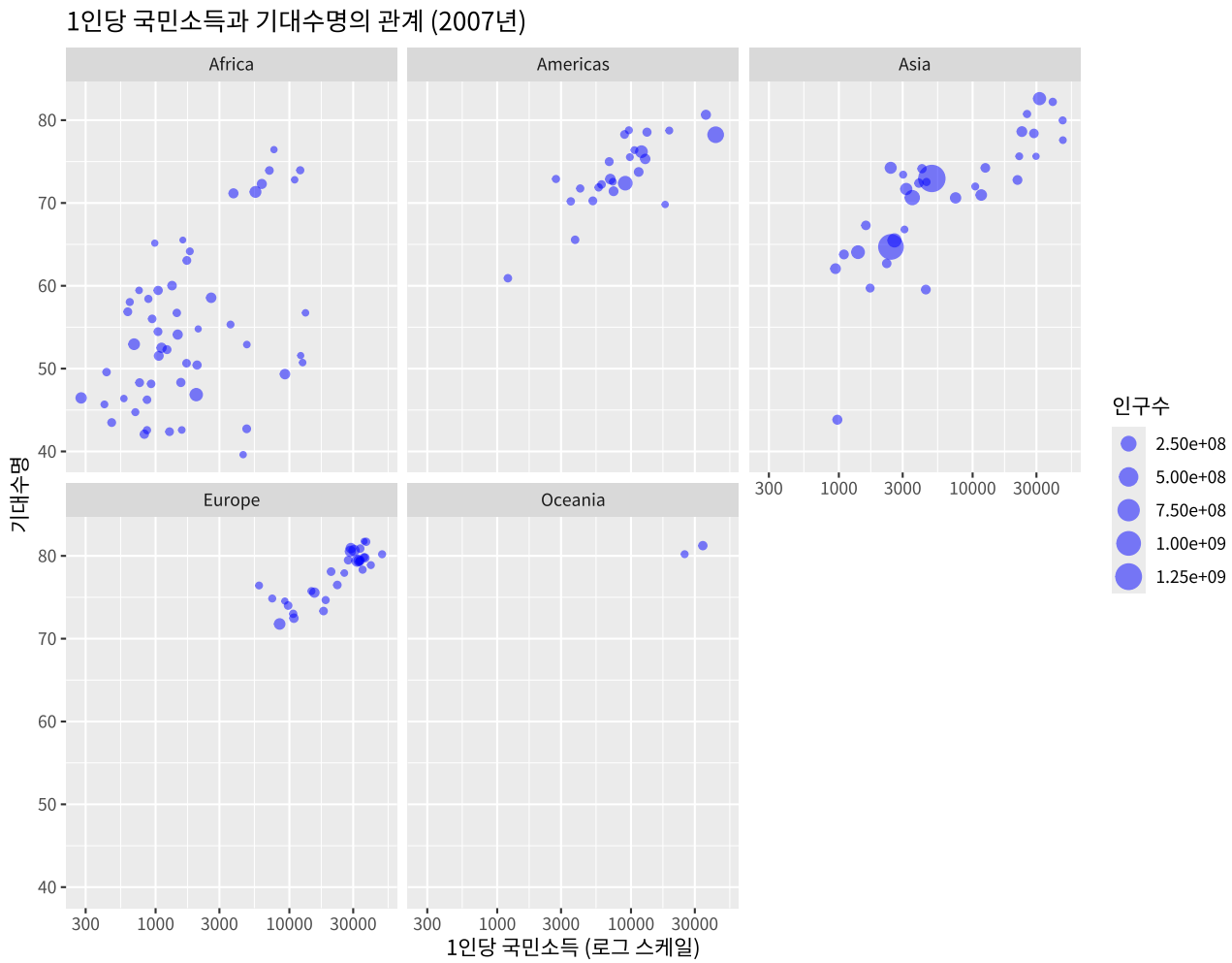
위의 그림에서 나라가 속한 대륙에 따라 점의 색깔을 다르게 나타내어 보자. 많은 경우 자료는 특정 그룹별로 분석하는 경우가 많기 때문에 그룹을 시각적으로 나타내는 것이 중요하다.

```
gapminder::gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, size=pop, color=continent)) +
  geom_point(alpha = 0.5) +
  scale_x_log10() + # x축을 로그 스케일로 변환
  labs(title = "1인당 국민소득과 기대수명의 관계 (2007년)",
       x = "1인당 국민소득 (로그 스케일)",
       y = "기대수명",
       size = "인구수",
       color = "대륙") +
  theme(text = element_text(family = "noto"))
```



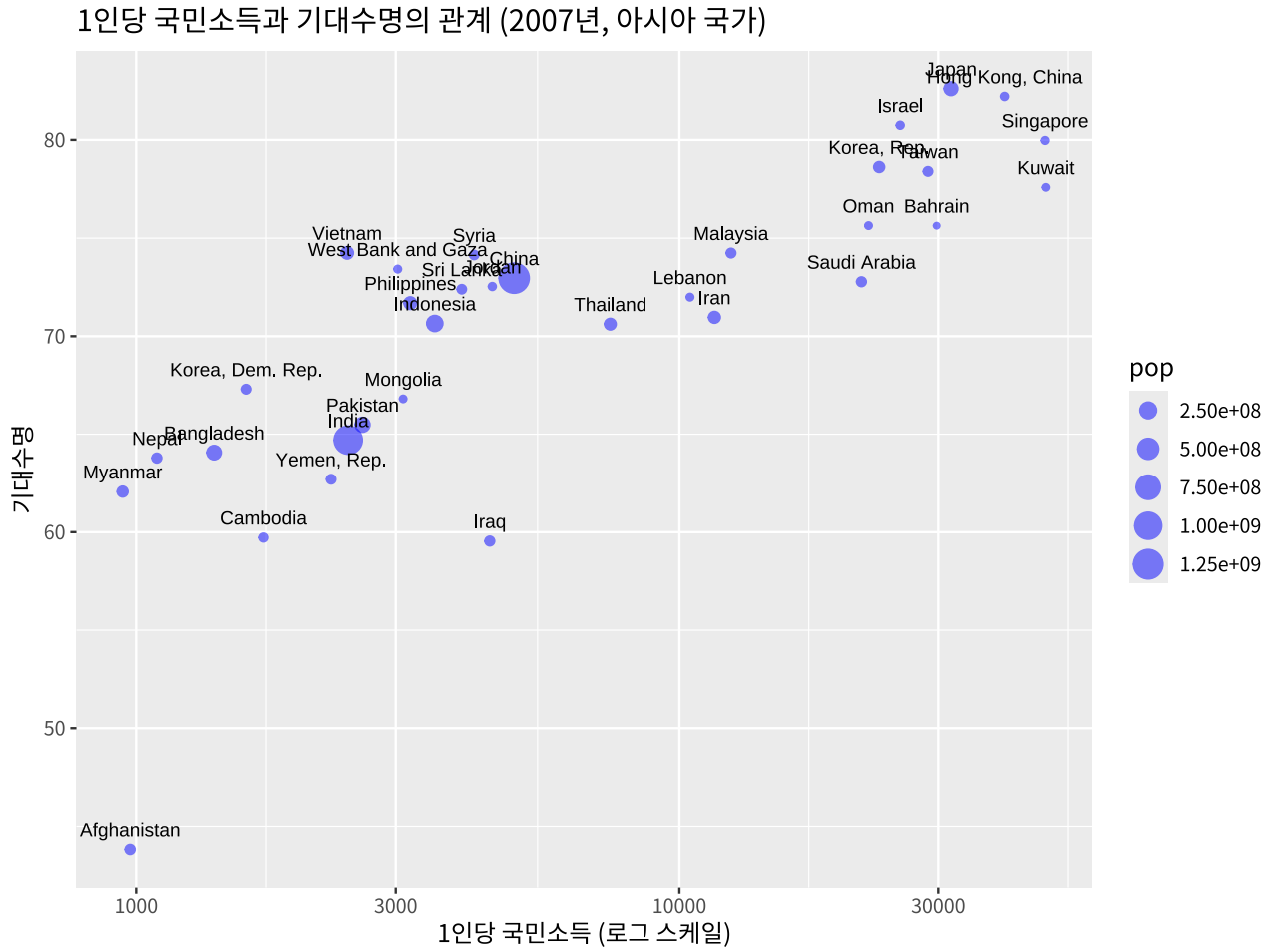
이렇게 그룹화하여 시각화 하는 경우 그룹마다 산점도를 따로 나타내는 방법으로 `facet` 기능을 사용할 수 있다. `facet` 기능은 `ggplot2` 패키지에서 제공하는 기능으로 `facet_wrap()` 함수와 `facet_grid()` 함수가 있다.

```
gapminder::gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, size=pop)) +
  geom_point(alpha = 0.5, color="blue") +
  scale_x_log10() +
  labs(title = "1인당 국민소득과 기대수명의 관계 (2007년)",
       x = "1인당 국민소득 (로그 스케일)",
       y = "기대수명",
       size = "인구수") +
  facet_wrap(~ continent) + # 대륙별로 산점도 따로 표시
  theme(text = element_text(family = "noto"))
```



위의 그림에서 아시아에 속한 나라들만 선택해서 점의 크기를 인구에 비례하게 하고 또한 나라의 이름을 표시해 보자.

```
gapminder::gapminder %>%
  filter(year == 2007, continent == "Asia") %>%
  ggplot(aes(x = gdpPercap, y = lifeExp, size=pop, label=country)) +
  geom_point(alpha = 0.5, color="blue") +
  geom_text(vjust = -1, size=3) + # 나라 이름 표시
  scale_x_log10() + # x축을 로그 스케일로 변환
  labs(title = "1인당 국민소득과 기대수명의 관계 (2007년, 아시아 국가)",
       x = "1인당 국민소득 (로그 스케일)",
       y = "기대수명",
       color = "대륙") +
  theme(text = element_text(family = "noto"))
```



2.2. 상관계수 행렬

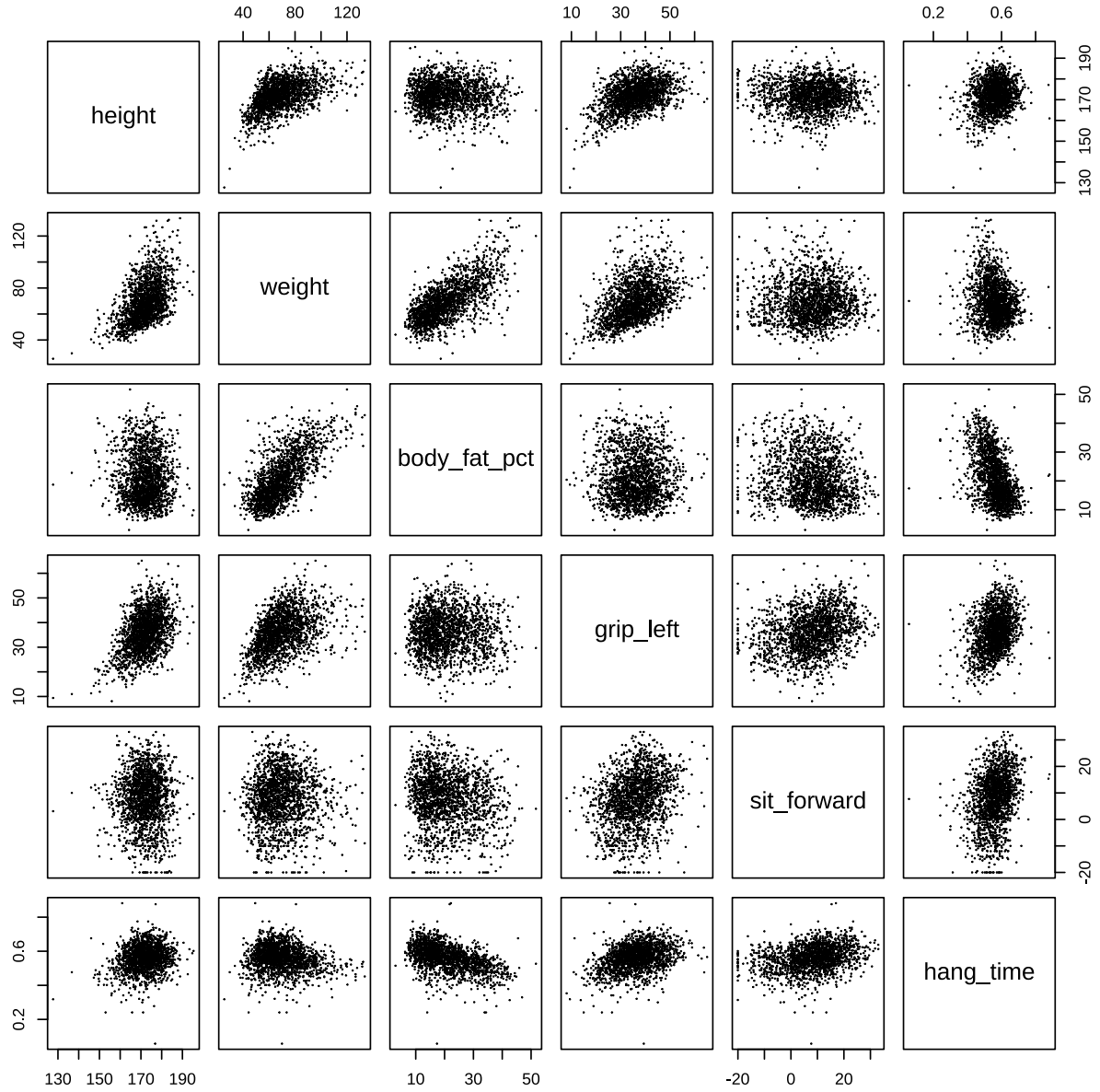
다변량 분석에서 상관계수는 여러 개의 변수들의 관계를 파악하는 가장 기본적인 통계량으로서 시각화를 이용하면 더 쉽게 이해할 수 있다.
상관계수 행렬(correlation matrix)은 여러 변수들 간의 상관계수를 한눈에 볼 수 있도록 정리한 그림이다. 상관계수 행렬을 시각화하는 방법으로는 히트맵(heatmap)이나 페어 플롯(pair plot) 등이 있다.

상관계수 행렬에 대한 예제는 앞 장에서 살펴본 국민체력100 자료를 이용하고자 한다.

먼저 가장 감단한 방법인 `pair()` 함수를 이용하여 페어 플롯을 그려보자. `pair()` 함수는 R에 기본으로 포함된 함수로서 여러 변수들 간의 산점도와 히스토그램을 한눈에 볼 수 있도록 그려준다.

```
load(here("data", "physical100.RData"))
# 국민체력100 자료에서 남자만 선택하여 데이터프레임 df 생성
df <- physical100_df %>%
  filter(sex == "남성") %>%
  select(height, weight, body_fat_pct, grip_left, sit_forward,hang_time)
```

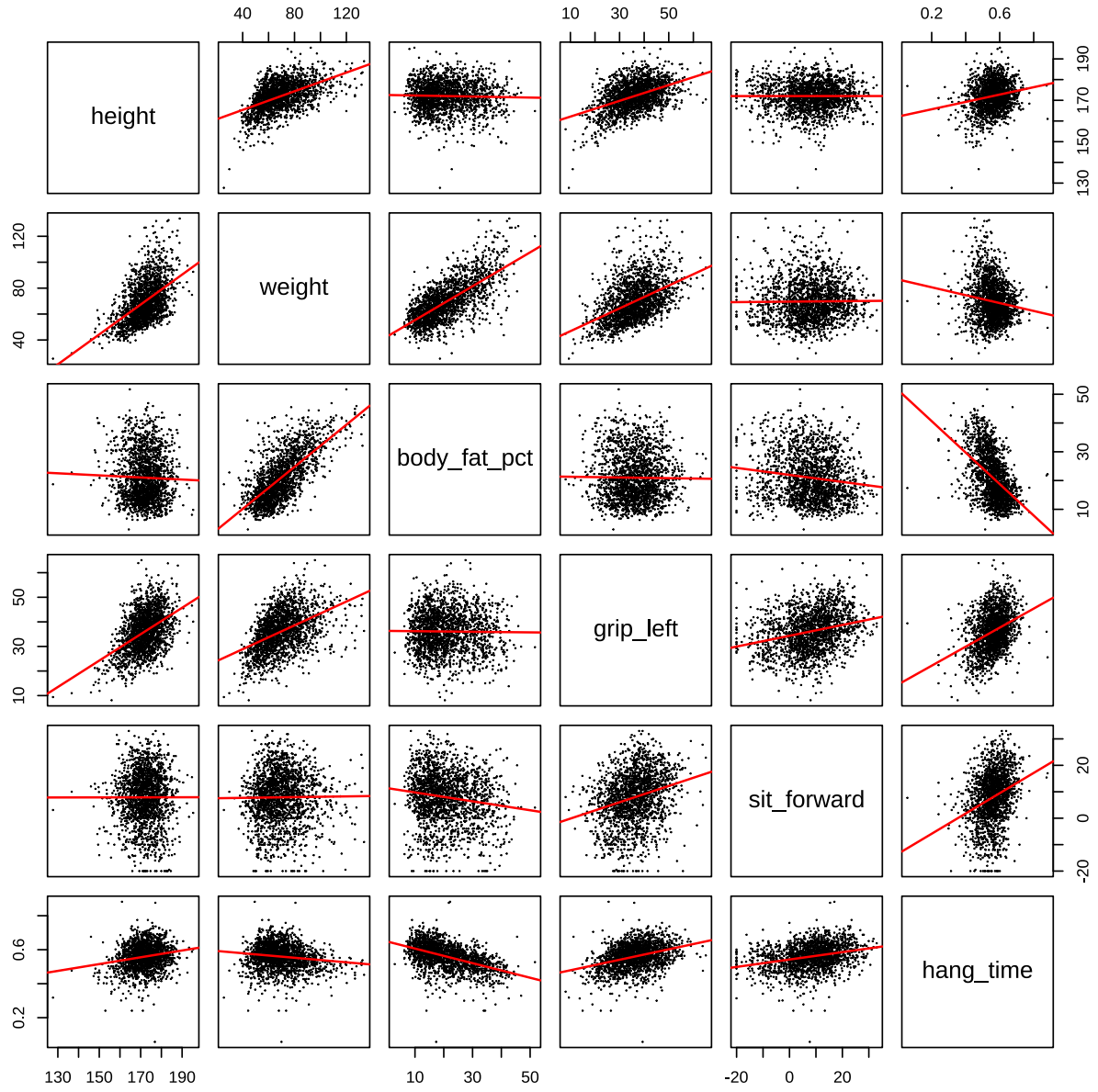
```
# pair() 함수를 이용하여 상관계수 행렬 그리기
pairs(df, pch=19, cex=0.1)
```



위와 같은 상관계수 산점도 행렬에 다음과 같이 회귀 직선을 추가하여 변수들 간의 관계를 더 명확히 나타낼 수 있다. 다음 코드는 `panel` 인수를 이용하여 각 산점도에 회귀 직선을 추가하는 방법을 보여준다.

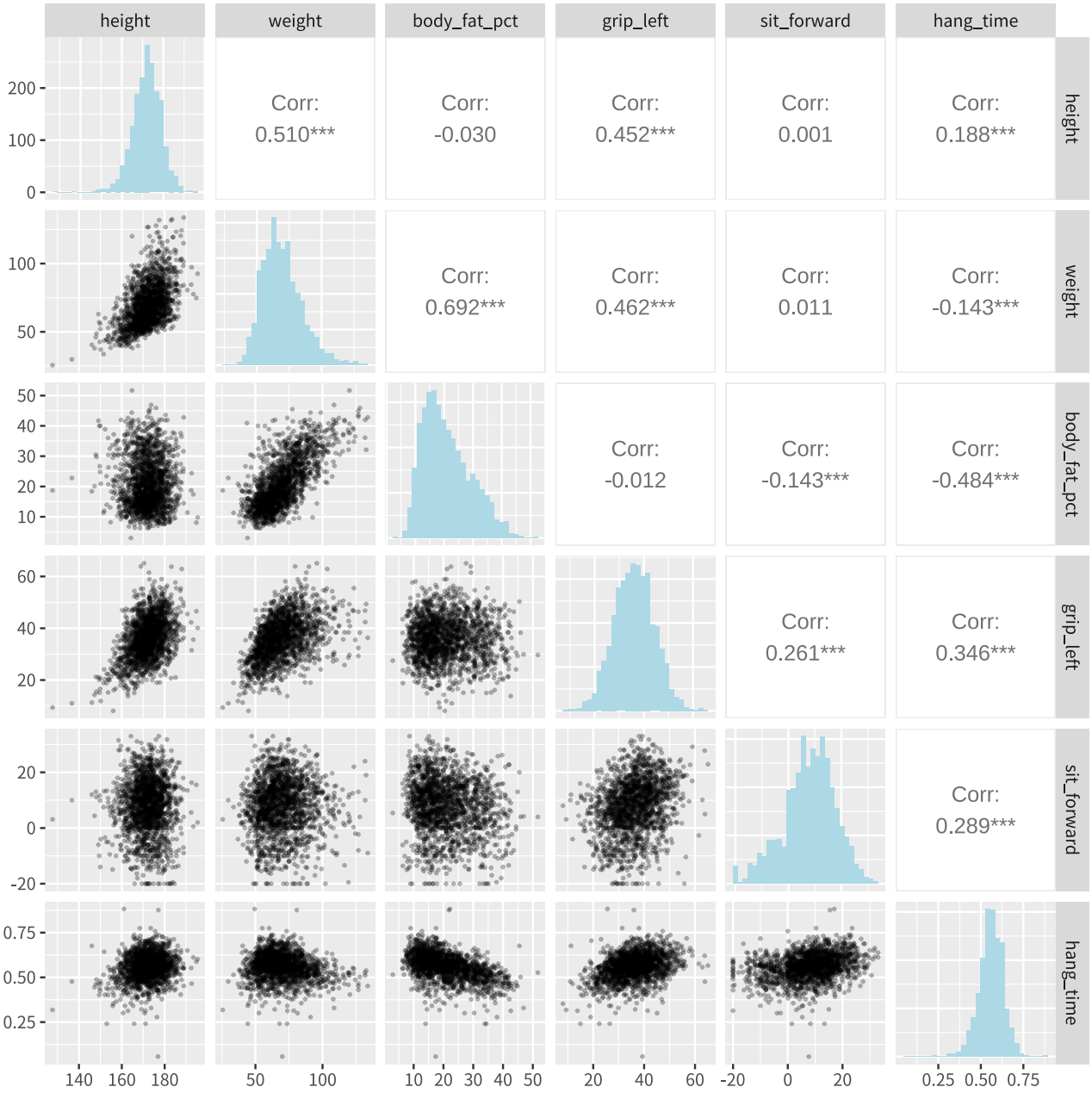
```
pairs(
  df,
  panel = function(x, y) {
    points(x, y, pch=19, cex=0.1)
    abline(lm(y ~ x), col = "red", lwd = 1.5)  # 선형회귀 직선
```

```
  }
)
```



위의 상관계수 행렬 그림에 좀 더 유용한 정보를 추가하는 다양한 방법들이 있다. 예를 들어, `GGally` 패키지의 `ggpairs()` 함수를 이용하면 상관계수 행렬에 상관계수 값과 각 변수에 대한 히스토그램을 추가하여 볼수 있다.

```
ggpairs(df,
        upper = list(continuous = wrap("cor", size = 4)), # 상관계수 표시
        lower = list(continuous = wrap("points", alpha=0.3, size=0.5)), # 산점도
        diag = list(continuous = wrap("barDiag", fill="lightblue"))) + # 히스토그램
  theme(text = element_text(family = "noto"))
```



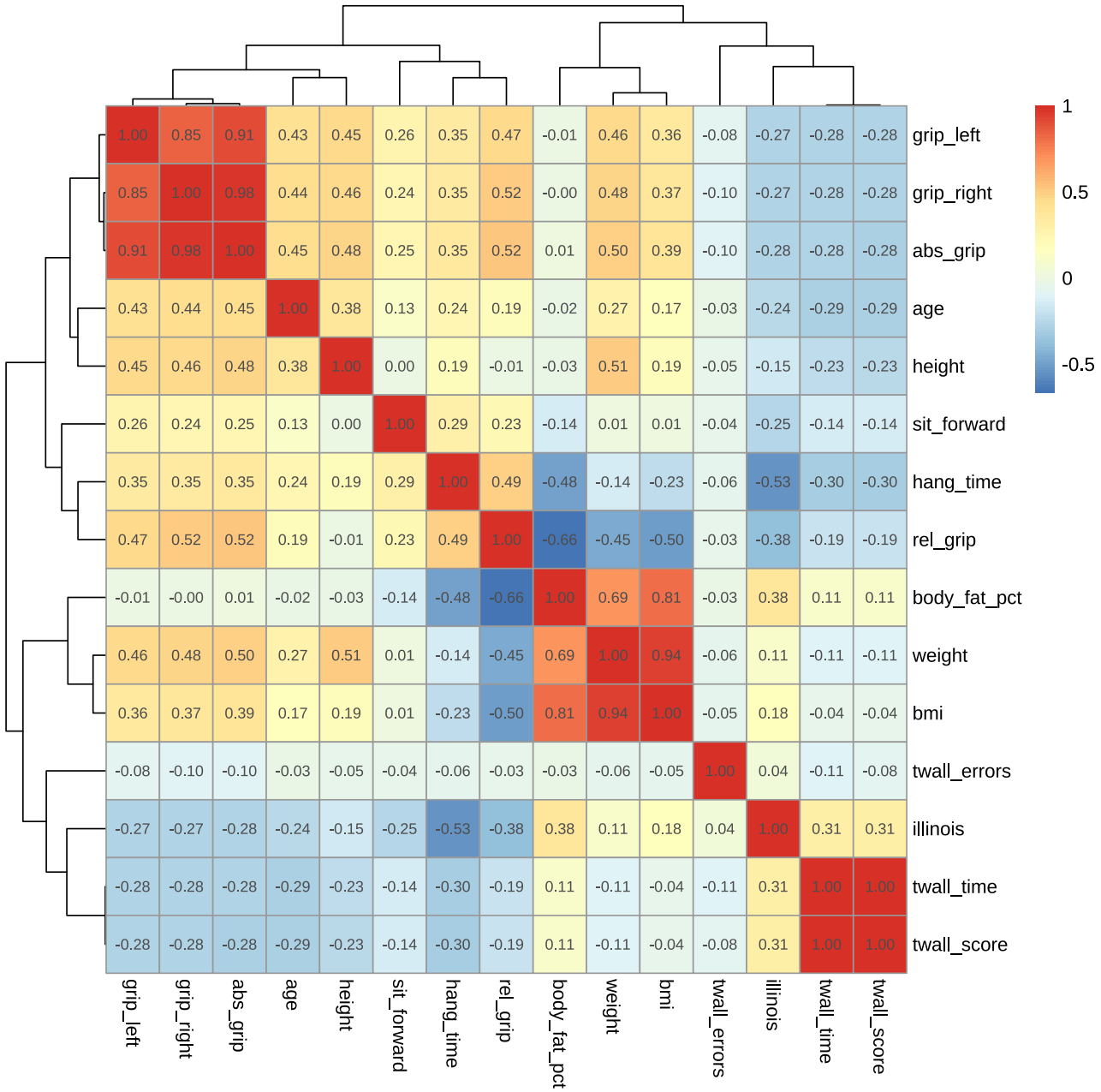
최근에는 상관계수 행렬을 히트맵(heatmap)으로 나타내기도 한다. 히트맵은 색깔로 값의 크기를 나타내는 그림으로서 상관계수 행렬을 히트맵으로 나타내면 변수들 간의 관계를 쉽게 파악할 수 있다.

특히, 변수의 개수가 매우 많은 경우 산점도를 이용한 상관계수 행렬은 시각적으로 분석이 어렵기 때문에 히트맵으로 나타내는 것이 더 유용할 수 있다. 패키지 `pheatmap` 의 `pheatmap()` 함수를 이용하여 상관계수 행렬을 히트맵으로 나타내보자. `pheatmap()` 함수의 유용한 점은 상관계수가 큰 변수끼리 군집화(clustering)하여 시각화할 수 있다는 점이다

자료는 국민체력100 자료에서 남자에 대한 모든 변수를 이용한 자료를 이용한다.

```
# 국민체력100 자료에서 남자만 선택하여 데이터프레임 df 생성
df <- physical100_df %>%
  filter(sex == "남성") %>%
```

```
  select(-sex)

# 상관계수 행렬 계산
cor_mat <- cor(df, use="pairwise.complete.obs") # 결측치가 있는 경우 pairwise로 계산

# 히트맵 그리기
pheatmap::pheatmap(cor_mat,  display_numbers = TRUE, number_format = "%.2f")
```



3. 다변량 가설 검정

```
library(Hotelling)
library(tidyverse)
library(here)
library(knitr)
library(purrr)
library(rmarkdown)
library(kableExtra)
library(flextable)
```

이번 장에서는 다변량 벡터의 자료에 대한 가설 검정법을 간단히 학습한다. 다변량에서 평균에 대한 검정은 일변량 벡터에서의 t-겁정의 개념을 확장하여 이해하는 것이 중요하다. 일변량 확률 분포에서 두 그룹의 평균을 비교하는 t-겁정 방법을 다변량으로 확장하는 방법을 단계별로 살펴보자.

참고로 이 장에서는 두 그룹에 대한 분산 또는 공분산이 같다고 가정한다. 공분산이 다른 경우는 아래 방법들을 확장해서 적용할 수 있지만 좀 더 복잡한 통계적 추론이 필요하다.

3.1. t-검정

기초통계학에서 나오는 가장 기본적이고 자주 쓰이는 가설검정 방법은 두 집단의 평균의 차이를 검정하는 t-검정(t-test)이다.

두 집단이 평균이 다르고 분산이 동일한 정규분포 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$를 따른다고 가정하고 다음과 같이 각각 n_1, n_2개의 독립 표본을 얻었다고 하자.

$$X_1, X_2, \dots, X_{n_1} \sim N(\mu_1, \sigma^2), \quad Y_1, Y_2, \dots, Y_{n_2} \sim N(\mu_2, \sigma^2)$$

위의 가설을 다음과 같은 t-통계량을 이용하여 검정할 수 있다.

$$t_0 = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{1/n_1 + 1/n_2}} \tag{3.1}$$

여기서 $\bar{X}, \bar{Y}$은 각 그룹의 표본 평균이다. 또한 S_p^2 은 두 집단의 공통분산 추정량(pooled variance estimator)이며 다음과 같이 계산한다.

$$\hat{\sigma}^2 = S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

식 3.1 의 t-검정 통계량의 분자는 집단 간의 평균의 차이를 나타낸다. 즉 $\bar{X} - \bar{Y}$는 두 집단의 표본 평균의 차이를 추정하는 양이고 그 차이가 크면 클수록 두 집단의 모평균의 차이 $\mu_1 - \mu_2$가 크다는 것을 의미한다.

t-검정 통계량의 분모는 두 집단의 공통분산 추정량 $\hat{\sigma}^2 = S_p^2$에 비례한다. 즉 집단 내의 변동을 반영하는 S_p^2 이 크면 클수록 t-검정 통계량은 그 크기가 작아져서 두 그룹 간에 차이가 있다는 증거가 약해진다.

정리해보면 t-검정 통계량은 **집단 간의 변동**(between-group variation)을 **집단 내의 변동**(within-group variation) 으로 나누어준 값이다. 다른 말로 급간 변동과 급내 변동을 사용하기도 한다.

이제 t-검정 통계량을 **제곱하면** 다음과 같이 표현할 수 있다.

$$t_0^2 = \frac{(\bar{X} - \bar{Y})^2}{S_p^2(1/n_1 + 1/n_2)} = \frac{\text{between-group variation}}{\text{within-group variation}} \tag{3.2}$$

통계학에서 등장하는 평균에 대한 겁정 통계량은 식 식 3.2 과 같이 그룹간 변동과 그룹내 변동의 비(ratio)로 이루어진 경우가 많다. 이제 더 나아가 검정 통계량의 다른 해석으로 통계적 거리의 의미를 살펴보자.

3.2. 통계적 거리

다변량 벡터의 평균에 대한 가설 검정을 위해서 다변량 벡터의 통계적 거리(statistical distance)의 개념에 대해서 알아보자.

먼저 일변량 벡터의 가설검정에 나타나는 t-검정 통계량의 형태를 식 3.2 으로 나타내면 다음과 같은 사실을 알 수 있다.

- 분자는 두 그룹의 평균에 대한 추정량의 공간적 거리, 즉 $\bar{X} - \bar{Y}$ 로서 두 그룹의 평균이 유클리디안 공간(Euclidean distance)에서 얼마나 떨어져 있는 가를 나타낸다. 검정 통계량에서는 거리의 제곱을 사용하였다.
- 분모는 두 그룹의 거리에 대한 통계적 불확실성을 반영한다. 이는 추정량의 분산 S_p^2 를 통해서 나타내며, S_p^2 가 커지면 공간적인 거리가 동일해도 통계적인 의미에서의 거리는 줄어드는 것이다.

이제 두 p-차원 확률 벡터 $\boldsymbol{X}$ 와 $\boldsymbol{Y}$에 대한 공간적 거리 $d(\boldsymbol{X}, \boldsymbol{Y})$ 는 다음과 같이 정의된다.

$$d(\boldsymbol{X}, \boldsymbol{Y})^2 = (\boldsymbol{X} - \boldsymbol{Y})^t(\boldsymbol{X} - \boldsymbol{Y})$$

이제 두 확률 벡터의 차이 $\boldsymbol{X} - \boldsymbol{Y}$ 의 불확실성을 나타내는 공분산 행렬을 $\boldsymbol{\Sigma}$ 하면, 확률 벡터의 통계적 거리(statistical distnace 또는 Mahalanobis distance) 는 다음과 같이 정의 된다.

$$d(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\Sigma})^2 = (\boldsymbol{X} - \boldsymbol{Y})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{Y}) \tag{3.3}$$

여기서 Σ^{-1} 은 공분산 행렬 Σ 의 역행렬(inverse matrix)이다.

식 3.3 의 통계적 거리는 일변량에서 사용되는 t-검정 통계량의 형태를 다변량 확률변수에 대해서 확장할 수 있는 방법을 제공해 준다.

3.3. 호텔링의 T^2 검정

이제 다변량 벡터의 평균에 대한 검정을 위해서 위에서 정의한 통계적 거리를 이용하여 호텔링의 T^2 검정(Hotelling's T^2 test)을 살펴보자.

확률 벡터 $\boldsymbol{X}$ 과 $\boldsymbol{Y}$ 가 평균이 각각 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 p-차원 다변량 정규 분포를 따른다고 가정하자.

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \boldsymbol{Y} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

호텔링의 T^2 검정은 두 그룹의 다변량 평균 벡터가 같은지에 대한 가설검정 방법이다. 즉 다음과 같은 가설을 검정한다.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

이제 가설 검정을 위하여 두 그룹에서 각각 n_1, n_2 개의 다변량 표본이 관측되었다고 하자.

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_{n_1} \sim_{IID} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_{n_2} \sim_{IID} N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

평균 벡터에 대한 추정량은 각 표본 평균 $\bar{\boldsymbol{X}}$ 과 $\bar{\boldsymbol{Y}}$ 이고 공분산 행렬의 합동 추정량 $\boldsymbol{S}_p$ 을 다음과 같이 정의한다.

$$\boldsymbol{S}_p = \hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^t + \sum_{i=1}^{n_2} (\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^t \right) \tag{3.4}$$

두 그룹의 평균벡터가 같은지에 대한 검정을 위하여 호텔링의 T^2 검정 통계량은 식 3.3 의 형태로 다음과 같이 정의된다.

$$\begin{aligned} T^2 &= (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^t \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \boldsymbol{S}_p \right]^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) \\ &= \frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^t \boldsymbol{S}_p^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) \end{aligned} \tag{3.5}$$

위의 호텔링의 T^2 통계량은 두 그룹의 평균 벡터 $\bar{\boldsymbol{X}}$ 와 $\bar{\boldsymbol{Y}}$ 의 차이에 대한 통계적 거리(statistical distance)를 나타낸다. 즉 두 그룹의 평균 벡터의 차이 $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$ 에 대한 제곱항을 그 차이의 불확실성을 나타내는 공분산 행렬 $\boldsymbol{S}$ 의 역행렬로 나누어준 값이다. 일변량에서 t-검정과 동일하게 식 3.5 의 값이 커지면 귀무가설에 반하는 정도가 커지는 것이다.

호텔링 통계량 T^2 은 귀무가설이 참인 경우, 즉 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ 일때 자유도가 각각 p 와 $n_2 + n_2 - p - 1$ 을 가지는 F-분포를 따른다. 이 때 p 는 확률 벡터의 차원이다.

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \sim F_{p, n_1+n_2-p-1} \quad \text{if } H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2} \text{ is true}$$

따라서 유의수준 α 에서 귀무가설을 검정하기 위해서는 다음과 같이 F-분포에서 기각역을 구하여 T^2 값과 비교한다.

$$\text{Reject } H_0 \quad \text{if } \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 > F_{\alpha; p, n_1+n_2-p-1}$$

또는 다음과 같이 계산한 p-값(p-value) 가 유의수준 α 보다 작으면 귀무가설을 기각한다.

$$\text{p-value } = P\left(F_{p, n_1+n_2-p-1} > \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \right)$$

3.4. 예제: 두 그룹의 평균벡터 검정

이 예제에서는 다변량 통계학에서 가장 자주 사용되는 피셔의 아이리스(Fisher's Iris) 자료를 이용하여 두 그룹의 평균벡터에 대한 검정을 배워보자.

R에 내장된 `iris`(Fisher's Iris) 자료는 1930년대 식물학자 Edgar Anderson 가 채집하고 측정한 붓꽃(iris) 데이터를 통계학자 R. A. Fisher(1936) 가 선형판별분석(Linear Discrimination Anslysis) 예제로 분석하면서 널리 알려졌다. 총 3개 종(Setosa, Versicolor, Virginica) 에서 각각 50개 표본으로 구성된 균형 자료(balanced data)이며 붓꽃의 특성을 나타내는 4개의 변수(단위: cm)로 구성되어 있다.

- `Sepal.Length`: 꽃받침 길이
- `Sepal.Width` : 꽃받침 너비
- `Petal.Length`: 꽃잎 길이
- `Petal.Width` : 꽃잎 너비
- `Species`: 범주형(세 종: `setosa`, `versicolor`, `virginica`)

```
data(iris)
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

이제 iris 자료에서 versicolor 와 virginica 두 종(각 50개 표본, $p = 4$ 변수)로 두 종에 대한 평균벡터가 동등한지 검정을 R 프로그램으로 수행해보자. 먼저 두 개의 종만 포함하는 자료를 만들고 표본 통계량을 구해보자.

```
# 패키지
#install.packages(c("Hotelling", "biotools"), dependencies = TRUE)
##library(Hotelling)
#library(biotools)

df <- iris %>%
      filter(Species %in% c("versicolor", "virginica"))

df$Species <- droplevels(df$Species)  # 두 수준만

head(df)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1          7.0         3.2          4.7         1.4 versicolor
2          6.4         3.2          4.5         1.5 versicolor
3          6.9         3.1          4.9         1.5 versicolor
4          5.5         2.3          4.0         1.3 versicolor
5          6.5         2.8          4.6         1.5 versicolor
6          5.7         2.8          4.5         1.3 versicolor
```

```
# 각 그룹의 표본 크기
n1 <- sum(df$Species=="versicolor")
n2 <- sum(df$Species=="virginica")
p <- ncol(df)-1  # 변수 개수
n1; n2; p
```

```
[1] 50
```

```
[1] 50
```

```
[1] 4
```

다음으로 두 그룹에 대한 평균 벡터와 두 그룹의 평균의 차이를 나타내는 벡터를 구해보자.

```
# 그룹별 4개의 변수에 대한 평균
mean_vec <- df %>% group_by(Species) %>%
       summarise(across(Sepal.Length:Petal.Width, list(mean=mean), .names="{col}_{fn}"))
mean_vec
```

```
# A tibble: 2 x 5
  Species  Sepal.Length_mean Sepal.Width_mean Petal.Length_mean Petal.Width_mean
  <fct>                <dbl>            <dbl>             <dbl>            <dbl>
1 versico~              5.94             2.77              4.26             1.33
2 virgini~              6.59             2.97              5.55             2.03
```

```
mean_x <- mean_vec %>%
  filter(Species=="versicolor") %>%
  select(Sepal.Length_mean, Sepal.Width_mean, Petal.Length_mean, Petal.Width_mean) %>%
  as.matrix()

mean_y <- mean_vec %>%
  filter(Species=="virginica") %>%
  select(Sepal.Length_mean, Sepal.Width_mean, Petal.Length_mean, Petal.Width_mean) %>%
  as.matrix()

mean_diff <- mean_x - mean_y
mean_diff
```

```
     Sepal.Length_mean Sepal.Width_mean Petal.Length_mean Petal.Width_mean
[1,]            -0.652           -0.204            -1.292             -0.7
```

이제 두 그룹에 대한 공분산 행렬을 구하고 합동 분산 추정량을 구해보자.

```
# 그룹별 4개의 변수에 대한 공분산 행렬을 List 형식으로 저장
cov_tbl <- df %>%
  group_by(Species) %>%
  summarise(cov = list(cov(across(where(is.numeric)))), .groups = "drop")
cov_tbl
```

```
# A tibble: 2 x 2
  Species    cov
  <fct>      <list>
1 versicolor <dbl [4 x 4]>
2 virginica  <dbl [4 x 4]>
```

```
# versicolor 종의 공분산 행렬 꺼내기 (마지막에 .[[1]] 은 앞의 개체에서 첫 번째 요소를 추출하는 명령)
cov_x <- cov_tbl %>% filter(Species == "versicolor") %>% pull(cov) %>% .[[1]]
cov_x
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.26643265  0.08518367   0.18289796  0.05577959
Sepal.Width    0.08518367  0.09846939   0.08265306  0.04120408
Petal.Length   0.18289796  0.08265306   0.22081633  0.07310204
Petal.Width    0.05577959  0.04120408   0.07310204  0.03910612
```

```
# virginica 종의 공분산 행렬 꺼내기
cov_y <- cov_tbl %>% filter(Species == "virginica") %>% pull(cov) %>% .[[1]]
cov_y
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.40434286  0.09376327   0.30328980  0.04909388
Sepal.Width    0.09376327  0.10400408   0.07137959  0.04762857
Petal.Length   0.30328980  0.07137959   0.30458776  0.04882449
Petal.Width    0.04909388  0.04762857   0.04882449  0.07543265
```

```
# 합동 공분산 행렬
Sp <- ((n1-1) * cov_x  + (n2-1) * cov_y  ) / (n1 + n2 - 2)
Sp
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.33538776  0.08947347   0.24309388  0.05243673
Sepal.Width    0.08947347  0.10123673   0.07701633  0.04441633
Petal.Length   0.24309388  0.07701633   0.26270204  0.06096327
Petal.Width    0.05243673  0.04441633   0.06096327  0.05726939
```

이제 다음과 같이 위에서 구한 표본 통계량을 이용하여 Hotelling T^2 을 다음과 같이 구할 수 있다.

```
# 평균벡터와 공분산 행렬의 차원 확인
dim(mean_diff); dim(Sp)
```

```
[1] 1 4
```

```
[1] 4 4
```

```
# Hotelling T^2 통계량 계산
T2 <- (n1*n2/(n1+n2)) * mean_diff %*% solve(Sp) %*% t(mean_diff)
T2
```

```
         [,1]
[1,] 355.4721
```

이제 기각역을 다음과 같이 구하고 위의 호텔링의 T^2 통계량과 비교해보자

```
# 유의수준
alpha <- 0.05
# F-분포의 임계값
F_crit <- qf(1-alpha, df1 = p, df2 = n1 + n2 - p - 1)
F_crit
```

```
[1] 2.467494
```

호텔링의 T^2 통계량의 값 355.4721452 이 기각역 2.4674936 보다 크므로 귀무가설을 기각한다. 즉, 두 종 versicolor와 virginica의 평균벡터가 통계적으로 유의하게 다르다고 할 수 있다.

위에서 직접구한 것과 동일한 결과를 주는 R 패키지 Hotelling의 hotelling.test() 함수를 사용하여 검정을 수행해보자. 이 함수는 등공분산 가정을 전제로 한다.

```
library(Hotelling)
res <- hotelling.test(Sepal.Length + Sepal.Width + Petal.Length + Petal.Width ~ Species, data
res
```

```
Test stat:  355.47
Numerator df:  4
Denominator df:  95
P-value:  0
```

4. 공분산 행렬의 추정

```
library(MVA)
library(mvtnorm)
library(gapminder)
library(GGally)
library(pheatmap)
library(mmrm)
library(nlme)

library(tidyverse)
library(here)
library(knitr)
library(purrr)
library(rmarkdown)
library(kableExtra)
library(flextable)
```

4.1. 공분산 행렬의 정의

공분산행렬은 다변량 분석에서 여러 변수들 간의 관계를 살펴볼 수 있는 가장 기본적이고 중요한 요소이다. 공분산 행렬의 구조를 이해하고 적절히 추정하는 것은 다변량 데이터 분석에서 매우 중요하다. 본 장에서는 공분산 행렬의 다양한 구조적 형태와 이를 추정하는 방법들을 간단히 소개한다.

먼저 p-차원 확률벡터 $\boldsymbol{X}$ 가 다변량 정규분포를 따른다고 가정하자.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

공분산 행렬 Σ 는 다음과 같이 정의되며

$$\boldsymbol{\Sigma} = \mathsf{Var}(\boldsymbol{X}) = \boldsymbol{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top], \quad \boldsymbol{X} \in \mathbb{R}^p$$

다음과 같은 성질을 가지고있다.

- 대각원소는 각 변수의 분산, 비대각원소는 변수 간의 공분산을 나타낸다.
- 대칭 행렬(symmeric matrix): $\mathbf{\Sigma}^\top = \mathbf{\Sigma}$
- 양반정치 행렬(semi-positive matrix):

$$\boldsymbol{a}^\top \mathbf{\Sigma} \boldsymbol{a} \geq 0 \quad \text{for all } \boldsymbol{a} \in \mathbb{R}^p$$

- 양반정치를 확인하는 방법은 공분산 행렬의 고유값이 모두 0 보다 같거나 커야한다.

이제 다변량 정규분포를 따르는 예제 데이터를 생성하고, 표본 공분산 행렬을 계산해 보자. 평균는 모두 0 이고 다음과 같은 공분산행렬을 가지는 6차원 다변량 정규분포를 고려한다. 다변량 정규분포에서 100개의 표본을 임의로 추출한 다음 표본 공분산 행렬을 구해보자.

$$\Sigma = \begin{bmatrix} 1.0 & 0.4 & 0.2 & 0.5 & 0.1 & -0.2 \\ 0.4 & 1.0 & 0.3 & 0.4 & -0.3 & 0.01 \\ 0.2 & 0.3 & 1.0 & 0.3 & 0.2 & -0.1 \\ 0.5 & 0.4 & 0.3 & 1.0 & 0.4 & -0.2 \\ 0.1 & -0.3 & 0.2 & 0.4 & 1.0 & 0.2 \\ -0.2 & 0.01 & -0.1 & -0.2 & 0.2 & 1.0 \end{bmatrix}$$

```r
# 예제 데이터 (다변량 정규)

# 공분산 행렬
Sigma_true <- matrix(c(
  1.0,  0.4,  0.2,  0.5,  0.1, -0.2,
  0.4,  1.0,  0.3,  0.4, -0.3,  0.01,
  0.2,  0.3,  1.0,  0.3,  0.2, -0.1,
  0.5,  0.4,  0.3,  1.0,  0.4, -0.2,
  0.1, -0.3,  0.2,  0.4,  1.0,  0.2,
 -0.2,  0.01, -0.1, -0.2,  0.2,  1.0
), 6, 6, byrow = TRUE)

Sigma_true
```

```
     [,1]  [,2] [,3] [,4] [,5]  [,6]
[1,]  1.0  0.40  0.2  0.5  0.1 -0.20
[2,]  0.4  1.00  0.3  0.4 -0.3  0.01
[3,]  0.2  0.30  1.0  0.3  0.2 -0.10
[4,]  0.5  0.40  0.3  1.0  0.4 -0.20
[5,]  0.1 -0.30  0.2  0.4  1.0  0.20
[6,] -0.2  0.01 -0.1 -0.2  0.2  1.00
```

```r
# 고유값 확인- 공분산행렬의 양반정치 점검
eigen(Sigma_true)$values
```

```
[1] 2.1507035 1.3633588 1.0108471 0.8176237 0.4860156 0.1714512
```

```r
# 다변량 정규분포에서 표본 생성
set.seed(123) # 재현 가능성
X <- rmvnorm(n = 100, mean = c(0, 0, 0, 0, 0, 0), sigma = Sigma_true)
colnames(X) <- c("x1", "x2", "x3", "x4", "x5", "x6")

# 표본의 일부 보기
head(X)
```

```
             x1           x2         x3          x4         x5         x6
[1,] -0.60697291 -0.016012988  1.3739779 -0.09555296  0.5483513  1.6489336
[2,]  0.06807651 -1.514672553 -0.7643606 -0.39888978  1.2909460  0.4933002
[3,]  0.98823956  0.285412440 -0.1300729  2.00665652  0.5701723 -2.0653381
[4,]  0.46989680 -0.333468735 -1.1591075 -0.42188577 -1.0883766 -0.8543093
[5,] -0.99065257 -1.222874588  0.3448429 -0.68099634 -0.4556374  1.0022332
[6,]  0.58628469 -0.005235656  1.0113511  1.05145508  1.2414359  0.5797083
```

```r
# 표본 공분산 행렬
S_hat <- cov(X)
S_hat
```

```
            x1          x2         x3         x4          x5          x6
x1  0.88057500  0.33490828  0.1872037  0.3207651 -0.05887987 -0.32665921
x2  0.33490828  0.88314376  0.3453332  0.2363243 -0.37813166 -0.06704301
x3  0.18720366  0.34533319  1.0006850  0.1587787  0.10083233 -0.01456920
x4  0.32076508  0.23632432  0.1587787  0.8100167  0.28207315 -0.44437452
x5 -0.05887987 -0.37813166  0.1008323  0.2820731  0.82579538  0.04660305
x6 -0.32665921 -0.06704301 -0.0145692 -0.4443745  0.04660305  1.06484897
```

```r
# 추정량과 실제 공분산 행렬 비교: 각 분산과 공분산의 상대적인 오차(%)
round(100 * (S_hat - Sigma_true) / (Sigma_true), 2)
```

```
        x1      x2     x3     x4      x5      x6
x1  -11.94  -16.27  -6.40 -35.85 -158.88   63.33
x2  -16.27  -11.69  15.11 -40.92   26.04 -770.43
x3   -6.40   15.11   0.07 -47.07  -49.58  -85.43
x4  -35.85  -40.92 -47.07 -19.00  -29.48  122.19
x5 -158.88   26.04 -49.58 -29.48  -17.42  -76.70
x6   63.33 -770.43 -85.43 122.19  -76.70    6.48
```

4.2. 공분산 행렬의 형태

- 일반형 공분산

p-차원 공분산 행렬의 일반적인 구조(general 또는 unstructured) 는 다음과 같으며 대칭 행렬이기 때문에 모수(parameter)의 개수는 $p(p-1)/2$ 개이다.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} \tag{4.1}$$

- 독립과 등분산

다변량 정규분포에서 모든 변수 $X_1, X_2, \dots, X_p$ 가 독립인 경우 공분산이 0인 것과 동일하기 때문에 다음과 같은 대각행렬의 형태를 가지게 된다.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$$

이 경우 다음과 같이 각 확률변수가 독립적으로 분산이 다른 일변량 정규분포를 따른다고 할 수 있다.

$$X_i \sim_{indep} N(\mu_i, \sigma_i^2), \quad i = 1, 2, \dots, p$$

더 나아가 모든 변수의 분산이 동일한 경우(등분산)에는 다음과 같은 구형(spherical) 형태의 공분산 행렬을 가진다.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}_p$$

- 균등 상관 구조

특별한 구조를 고려하는 경우, 공분산 행렬의 형태로 가장 자주 사용되는 구조가 균등 상관 구조(Compound Symmetry, CS) 이다. 이 형태는 분산이 모두 동일하고 공분산도 모두 동일한 형태이며 분산과 공분산은 다르게 설정된다. 즉 모든 변수의 분산이 σ^2 이고 모든 변수 쌍들의 공분산이 $\rho\sigma^2$ 인 경우이다. 따라서 ρ 는 상관 계수이다. 따라서 모든 변수의 상관계수도 동일한 형태를 가지는 구조이다.

$$cor(X_i, X_j) = \rho \quad \text{for all } i, j$$

$$\boldsymbol{\Sigma}_{\text{CS}} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\big[(1-\rho)\boldsymbol{I}_p + \rho\boldsymbol{J}_p\big] \tag{4.2}$$

위의 식에서 $\boldsymbol{J}_p$ 는 모든 원소가 1인 $p \times p$ 행렬이다.

4.2.1. AR(1) 구조

공분산 행렬의 또 다른 구조적 형태로는 AR(1) 구조가 있다. 이 형태는 시계열 자료에서 자주 사용되는 구조로서, 인접한 변수들 간의 상관관계가 멀어질수록 지수적으로 감소하는 특징을 가진다.

$$cor(X_i, X_j) = \rho^{|i-j|}$$

AR(1) 구조의 공분산 행렬은 다음과 같은 형태를 가진다.

$$\boldsymbol{\Sigma}_{AR} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \cdots & \rho^{p-1}\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \cdots & \rho^{p-2}\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & \cdots & \rho^{p-3}\sigma^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1}\sigma^2 & \rho^{p-2}\sigma^2 & \rho^{p-3}\sigma^2 & \cdots & \sigma^2 \end{bmatrix}$$

따라서 AR(1) 구조를 가진 다변량 확률 벡터의 상관 계수 행렬은 다음과 같이 나타낼 수 있다.

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{bmatrix}$$

4.2.2. 블록 대각 구조

공분산 행렬의 또 다른 구조적 형태로는 블록 대각(Block Diagonal) 구조가 있다. 이 형태는 변수들이 여러 개의 그룹으로 나누어져 있고, 각 그룹 내에서는 변수들 간에 상관관계가 존재하지만, 그룹 간에는 상관관계가 없는 경우에 적합하다. 블록 대각 구조의 공분산 행렬은 다음과 같은 형태를 가진다.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_3 \end{bmatrix}$$

4.3. 공분산의 추정

4.3.1. 표본 공분산 행렬

만약 n 개의 표본 $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ 이 관측되었다고 하자.

만약 분포의 가정에서 공분산이 제약이 없는 일반적인 형태 식 4.1 라고 한다면 다음과 같은 표본 공분산 행렬을 이용하여 추정한다. 이 추정량은 불편추정량(unbiased estimator)이고 동시에 n 이 충분히 크면 최대가능도 추정량과 동일하다고 볼 수 있다.

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \tag{4.3}$$

특별한 구조를 가진 공분산 행렬은 최대가능도 추정을 이용하여 추정할 수 있다.

4.3.2. 예제: 반복측정자료

이 장에서는 의학통계에 자주 사용되는 반복측정자료 형태의 다변량 확률벡터에 대하여 균등 상관 구조 형태의 공분산 행렬을 추정하는 예제에 대해서 살펴 보자.

확률변수 X_i 는 i 시점에서 순서대로 5번 관측한 자료이며 한 명의 개체가 각 시점에서 반응값을 측정한다고 하자. 따라서 반복으로 측정한 확률 변수 $X_1, X_2, \dots, X_5$ 는 독립이 아니다.

이제 확률 벡터 $\boldsymbol{X} = (X_1, X_2, \dots, X_5)^t$ 가 다변량 정규분포를 따른다고 가정하자. 5-차원의 다변량 정규 확률벡터를 고려하고 각 변수의 평균은 시간을 나타내는 시점 (i)에 비례하게 다음과 같이 정의한다.

$$\mu_i = E(X_i) = 0.5 + 0.1(i-1)$$

공분산 행렬은 식 4.2 의 균등 상관 구조를 가지며 분산은 모두 1 이고 상관계수는 0.6으로 가정한다. 표본의 개수는 100 개이다.

먼저 주어진 평균벡터와 공분산 행렬에 대하여 분포를 정의하고 가상의 자료를 임의로 추출하는 R 코드를 고려한다.

```
set.seed(121)
n  <- 100          # 표본 수
p  <- 5            # 확률 벡터의 차원

# 각 개체의 id 와 시간 변수 생성
id <- factor(rep(1:n, each = p))
time <- rep(1:p, times = n)
mu <- 0.5 + 0.1 * (rep(1:p, times = 1) - 1)  # 평균 벡터

#   p- 차원 CS 공분산을 만드는 함수
```

```
make_Sigma_CS <- function(p, sigma2 = 1, rho = 0.3) {
  if (rho <= -1/(p - 1) || rho >= 1) {
    stop("rho must be in (-1/(p-1), 1) to ensure positive definiteness.")
  }
  J <- matrix(1, p, p)          # 모든 원소가 1인 행렬
  Sigma <- sigma2 * ((1 - rho) * diag(p) + rho * J)
  return(Sigma)
}

# 균등 상관 구조의 공분산 행렬
Sigma_true <- make_Sigma_CS(p, sigma2 = 1, rho = 0.6)
Sigma_true
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]  1.0  0.6  0.6  0.6  0.6
[2,]  0.6  1.0  0.6  0.6  0.6
[3,]  0.6  0.6  1.0  0.6  0.6
[4,]  0.6  0.6  0.6  1.0  0.6
[5,]  0.6  0.6  0.6  0.6  1.0
```

이제 100 개의 표본을 다변량 정규분포에서 임의로 추출하고 wide 형식의 자료로 변환해 보자.

```
Xmat <- rmvnorm(n = n, mean = mu, sigma = Sigma_true) # 다변량 표본 추출
colnames(Xmat) <- paste0("X", 1:p)

df_wide <- as.data.frame(Xmat) # wide 형식의 자료
head(df_wide)
```

```
          X1          X2         X3          X4          X5
1  0.1412804  0.47132670  0.5836000  0.55150916  0.25141129
2  1.9259813  1.19236353  1.0678925  0.50890233  1.88213477
3  1.4160815  0.08403608  0.4753524  0.47093252  1.27335173
4  1.1549582 -0.30705522  1.4909395  0.08919076  1.23971146
5  1.2631167  1.31918169  1.2023846  1.90247109  1.83556681
6 -0.6336252  0.47159709 -0.6432958 -0.37359215 -0.06331619
```

추출된 표본을 이용하여 식 4.3 에 주어진 표본 공분산 행렬을 계산해 보자. 표본 공분산 행렬은 제약조건이 없는 형태로 나타나기 때문에 모든 분산과 공분산이 각각 다르게 추정된다.

```
cov(df_wide)
```

```
          X1        X2        X3        X4        X5
X1 0.7180133 0.4216849 0.3886342 0.3553769 0.4661159
X2 0.4216849 0.8704348 0.4896497 0.4478777 0.4441199
X3 0.3886342 0.4896497 0.8766630 0.4369850 0.5013652
X4 0.3553769 0.4478777 0.4369850 0.7887407 0.4043893
X5 0.4661159 0.4441199 0.5013652 0.4043893 0.9236021
```

이제 균등 상관 구조를 가지는 공분산을 추정할 수 있는 방법을 알아보자. 먼저 자료는 측정 시간을 변수로 하는 긴 형식의 자료로 생성한다.

```r
# pivot_longer 함수를 이용하여 긴 형식의 자료를 생성
df_long <- df_wide %>%
  mutate(id = factor(1:n)) %>%
  pivot_longer(cols = starts_with("X"),
               names_to = "time",
               values_to = "y",
               names_prefix = "X") %>%
  mutate(time = as.integer(time))
head(df_long, 10)
```

```
# A tibble: 10 x 3
   id     time     y
   <fct> <int> <dbl>
 1 1         1 0.141
 2 1         2 0.471
 3 1         3 0.584
 4 1         4 0.552
 5 1         5 0.251
 6 2         1 1.93
 7 2         2 1.19
 8 2         3 1.07
 9 2         4 0.509
10 2         5 1.88
```

공분산 추정을 위하여 `nlme` 패키지에 있는 다변량 확률변수의 회귀식을 추정하는 함수 `nlme` 를 사용하여고 하며, 균등 상관 구조를 가지는 공분산을 가정한다.

다음 추정된 다변량 회귀모형의 결과를 보면 식 4.2 의 균등 상관 구조에서 분산 σ 의 추정값은 $0.8362 = (0.9144)^2$, 상관계수 ρ 의 추정값은 0.52078 로 나타난다. 참고로 아래 결과에서 `rho` 는 공통 상관계수의 추정값,`Residual standard error` 는 표준편차 추정값이다.

```r
fit_gls_cs <- gls(y ~ time, data = df_long,
                  correlation = corCompSymm(form = ~ 1 | id))

summary(fit_gls_cs)
```

```
Generalized least squares fit by REML
  Model: y ~ time
  Data: df_long
       AIC      BIC    logLik
  1163.245 1180.088 -577.6226

Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.5207752

Coefficients:
                Value  Std.Error  t-value p-value
(Intercept) 0.3750164 0.09360729 4.006274   1e-04
time        0.0817480 0.02001772 4.083779   1e-04

 Correlation:
     (Intr)
time -0.642

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.76158379 -0.71458351  0.04453439  0.76834584  3.26193595

Residual standard error: 0.9144187
Degrees of freedom: 500 total; 498 residual
```

```r
# rho 는 공통 상관계수, Residual standard error 는 표준편차 추정값
```

위에서 주어진 결과를 가지고 공분산 행렬의 추정값을 다음과 같이 구할 수 있다.

```r
getVarCov(fit_gls_cs)
```

```
Marginal variance covariance matrix
        [,1]     [,2]     [,3]     [,4]     [,5]
```

```
[1,] 0.83616 0.43545 0.43545 0.43545 0.43545
[2,] 0.43545 0.83616 0.43545 0.43545 0.43545
[3,] 0.43545 0.43545 0.83616 0.43545 0.43545
[4,] 0.43545 0.43545 0.43545 0.83616 0.43545
[5,] 0.43545 0.43545 0.43545 0.43545 0.83616
  Standard Deviations: 0.91442 0.91442 0.91442 0.91442 0.91442
```

5. 판별분석

```
library(mvtnorm)
library(MASS)

library(tidyverse)
library(here)
library(knitr)
library(rmarkdown)
library(kableExtra)
library(flextable)

#아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
#font_add_google("Nanum Pen Script", "gl")
font_add_google(name = "Noto Sans KR", family = "noto")
showtext_auto()
```

의사결정은 주어진 유한개의 선택들 중에서 하나를 고르는 것이다. 예를 들어 은행에서 대출 신청자에게 돈을 빌려줄지 말지, 이동 시에 택시를 탈지, 버스를 탈지 또는 지하철을 탈지 결정해야 한다. 인간은 의사결정을 할 때 어느 정도의 자신만의 규칙에 따라 움직이며(아닌 경우도 많지만..) 여러 번의 시행 착오 등을 거쳐서 좋은 선택을 위해 규칙을 바꾸기도 한다. 또한 유한개의 선택들은 대체로 두 가지의 선택이 있으며 둘 중 하나를 선택하는 경우가 많다.

예를 들어, 신용평가에서 은행은 과거 경험을 통해 두 부류의 고객이 있음을 알고 있다. 즉, 대출을 아무 문제 없이 상환하는 안전 고객과 상환에 어려움을 겪은 위험 고객이다. 새로운 고객이 대출을 신청할 때, 은행은 대출을 해줄지 말지를 결정해야 한다. 은행의 과거 기록은 두 부류의 고객에 대한 다수의 특성값들를 제공하며, 여기에는 나이, 급여, 혼인 여부, 대출 금액 등과 같은 다양한 정보가 포함된다. 새로운 고객은 새로운 특성값을 가지고 있으며 판별규칙(Discrimination rule)은 이 새로운 고객을 두 집단 중 하나로 분류해야 하는 규칙을 말하는 것이다.

판별분석(Discriminant analysis)은 이러한 의사결정에서 통계적 분포과 방법을 사용하여 자료에 기반한 규칙을 정하는 방법이다. 기본적으로 전체 집단(모집단)이 두 개 이상의 집단들로 나누어져 있다고 생각하고 그 집단들에 대한 확률적 가정(분포 가정)을 고려한다. 같은 집단에 속하는 개체들은 유사하며 다른 집단에 속한 개체들은 그 특성이 다르다고 가정하고 이러한 집단의 특성을 확률적 분포로 나타낸다.

예를 들어 이동시 자가용을 타는 사람들과 대중교통을 타는 사람들은 소득의 분포가 다르다고 가정할 수 있다. 이러한 가정 아래 새로운 개체를 어느 집단에 배정하는지에 대한 규칙을 자료를 이용하여 정하는 방법이 판별분석

이며 다양한 분류방법(classification)의 출발점이다. 유의할 점은 규칙을 정할 때 어떤 규칙이 좋은 것인지에 대한 기준을 생각해야 한다. 동일한 상황에서 두 가지 규칙을 비교할 수 있어야 더 나은 의사결정을 할 수 있다.

판별분석에서는 다음과 같은 개념들을 생각해야 한다.

- 모집단을 구성하는 겹치지 않는 집단들(groups)
- 집단을 이루는 개체들의 특성(확률변수)과 그에 대한 분포
- 개체의 특성을 알 때 소속 집단을 결정하는 규칙
- 잘못된 판단(error)과 그에 상응하는 비용(cost)
- 어떤 규칙이 좋은 것인지 판단하는 기준

5.1. 분포와 판별규칙

이 장에서는 두 개의 집단(population) P_1 과 P_2 을 고려한다. 예를 들어 은행에서는 전체 고객을 안전 고객과 위험 고객, 두 집단으로 나눌 수 있다.

각 집단에 속하는 개체들의 특성(확률벡터) $\boldsymbol{X}$ 은 속하는 집단에 따라서 각각 분포 $F_i(\boldsymbol{x})$, $i = 1, 2$ 를 따른다고 가정한다. 또한 $f_i(\boldsymbol{x})$를 분포 F_i의 확률밀도함수라고 하자.

판별 규칙(discrimination rule) 은 새로운 개체의 특성 $\boldsymbol{X} = \boldsymbol{x}$ 가 주어졌을 때 이 개체가 어느 집단에 속하는지 결정하는 방법이다. 이러한 규칙은 개체의 특성값이 가질 수 있는 전체 공간(표본공간)을 겹치지 않는 두 집합 R_1 과 R_2 로 나누고 관측된 값이 R_1 에 속하면 새로운 개체를 P_1 에 배정하고, 반대로 R_2 에 속하면 P_2에 배정하는 것으로 주어진다.

$$\text{where } R_1 \cup R_2 = \mathbb{R}^p, \quad R_1 \cap R_2 = \emptyset \tag{5.1}$$

즉, R_1 과 R_2 는 전체 공간을 겹치지 않게 나누는 두 집합이다. 판별 규칙은 다음과 같이 쓸 수 있다.

$$\boldsymbol{x} \in R_1 \Rightarrow \text{개체를 } P_1\text{에 배정} \boldsymbol{x} \in R_2 \Rightarrow \text{개체를 } P_2\text{에 배정} \tag{5.2}$$

5.1.1. 판별 오류와 비용

이제 i 집단에 속한 개체가 주어진 판별 규칙에 따라서 j 집단에 배정될 사건과 확률을 생각해 보자. 먼저 i 집단에 속한 개체가 j 집단에 배정될 사건을 $A(j|i) \equiv A_{ji}$ 라고 정의한다. 따라서 각 개체가 자신이 속한 집단에 배정된 사건, A_{11} 과 A_{22} 가 발생하면 판별규칙이 잘 적용된 경우이다. 반대로 개체가 자신이 속하지 않은 다른 집단에 배정된 사건, A_{12} 과 A_{21} 가 발생하면 오류가 발생한 경우이며 이러한 경우 오류에 의한 비용 c_{12}, c_{21} 이 발생한다.

예를 들어 은행에서 대출 신청자가 우량 고객 P_1 에 속하는데 불량 고객 P_2 로 잘못 판별되어 대출을 거절하는 경우, 은행은 대출 이익을 얻지 못하는 손실 c_{21} 이 발생한다. 반대로 불량 고객 P_2 가 우량 고객 P_1 으로 잘못 판별되어 대출을 해주는 경우, 은행은 대출금을 상환받지 못하는 더 큰 손실 c_{12} 이 발생한다. 따라서 두 가지 오류에 대한 비용은 일반적으로 다르며, 이러한 비용을 고려하여 판별 규칙을 정하는 것이 중요하다.

5.1.2. 최대가능도 규칙

이제 위에서 정의한 사건과 비용을 이용하여 판별 분석에서 다루는 중요한 확률을 생각해보자. 먼저 첫 집단 P_1 에 속한 개체가 주어진 판별 규칙에 따라서 두 번째 집단 P_2 에 배정될 사건확률은 다음과 같다.

$$p_{21} \equiv P(A_{21}) = \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} \tag{5.3}$$

마찬가지로 P_2 에 속한 개체가 P_1 에 배정될 확률은 다음과 같다.

$$p_{12} \equiv P(A_{12}) = \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \tag{5.4}$$

위에서 정의한 두 확률 p_{12} 와 p_{21} 은 판별에서 오류가 발생한 확률이다.

판별 규칙을 만드는 것은 식 5.1 과 식 5.2 에서 정의한 두 집합 R_1 과 R_2 를 정하는 것이다. 따라서 잘못된 분류를 범할 확률 식 5.3 과 식 5.4 은 판별 규칙에 따라서 달라진다. 즉, R_1 과 R_2 를 다르게 정하면 오류 확률도 달라진다.

판별 규칙은 당연히 두 오류의 확률을 가장 작게 만드는 방향으로 정해져야 한다. 이제 앞에서 정의한 두 그룹의 분포를 이용하여 오류의 확률을 가장 작게 만드는 판별 규칙을 어떻게 정할 수 있는지 알아보자. 다음과 같이 두 오류를 범할 확률을 최소로 하는 판별 규칙을 찾으려 한다

$$\min_{R_1, R_2} \{p_{12} + p_{21}\} = \min_{R_1, R_2} \left[\int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} \right] \tag{5.5}$$

위의 식은 두 오류 확률의 합을 최소로 하는 판별 규칙을 찾는 문제이다. 두 종류의 오류를 범할 확률의 합을 다시 살펴보면 다음과 같이 쓸 수 있다.

$$\begin{aligned} p_{12} + p_{21} &= \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} + 1 - \int_{R_1} f_1(\boldsymbol{x}) d\boldsymbol{x} \\ &= 1 + \int_{R_1} \{f_2(\boldsymbol{x}) - f_1(\boldsymbol{x})\} d\boldsymbol{x} \end{aligned} \tag{5.6}$$

참고로 위의 식은 전체 공간에서의 확률밀도함수의 적분이 1임을 이용하였다.

$$\int_{R_1} f_1 \boldsymbol{x} d\boldsymbol{x} + \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} = 1$$

이제 식 5.6 의 적분값을 최소로 하는 R_1 을 찾으면 된다. 이 적분값을 최소로 하기 위해서는 적분하는 영역 R_1 에서 적분하는 함수 $f_2(\boldsymbol{x}) - f_1(\boldsymbol{x})$ 의 값이 음수인 부분만 포함되도록 하면 된다. 즉, 다음과 같은 조건을 만족하는 R_1 을 찾으면 된다.

$$R_1 = \{\boldsymbol{x}|f_2(\boldsymbol{x}) - f_1(\boldsymbol{x}) < 0\} = \{\boldsymbol{x}|f_1(\boldsymbol{x}) > f_2(\boldsymbol{x})\} \tag{5.7}$$

만약 R_1 을 식 5.7 과 같이 정하면 R_2 는 자동으로 다음과 같이 정해진다.

$$R_2 = \{\boldsymbol{x}|f_2(\boldsymbol{x}) \geq f_1(\boldsymbol{x})\} \tag{5.8}$$

위의 유도에서 판별 규칙을 만들 때 중요한 요인은 두 확률밀도함수의 비(ratio)이다. 두 확률밀도함수의 비가 1 보다 큰 값을 가지는 표본 $\boldsymbol{x}$ 의 영역과 1 보다 작은 값을 가지는 표본 $\boldsymbol{x}$ 의 영역이 각각 R_1 과 R_2가 되는 것이다.

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}$$

앞에서 살펴보았듯이, 판별 함수를 만드는 가장 기본적인 방법은 최대 가능도 규칙(Maximum Likelihood Rule)이다. 이 방법은 새로운 개체의 특성 $\boldsymbol{X} = \boldsymbol{x}$ 가 주어졌을 때, 각 집단에서 이 특성이 관측될 가능도(likelihood function)를 계산하여 더 큰 집단에 개체를 배정하는 방법이다. 즉, 다음과 같이 판별 규칙을 정의한다.

$$\boldsymbol{x} \in R_1 \text{ if } f_1(\boldsymbol{x}) > f_2(\boldsymbol{x}), \quad \text{and} \quad \boldsymbol{x} \in R_2 \text{ if } f_1(\boldsymbol{x}) \leq f_2(\boldsymbol{x})$$

5.1.3. 베이지안 규칙

만약 각 집단에 속할 사전 확률(prior probability)이 각각 π_1 , π_2 로 주어졌다고 하자. 사전확률은 관측하는 확률벡터의 분포를 고려하기 전에, 어떤 개체가 각 집단에 속할 가능성을 확률을 나타낸 것이다.

예를 들어 앞에서 생각한 예제에서 은행의 대출 신청자가 우량 고객 집단에 속할 가능성과 물량 고객 집단에 속할 가능성이 다를 수 있다. 대부분의 사람들이 대출을 계획대로 상환하였다면 임의의 고객이 우량 고객일 가능성이 불량 고객일 가능성보다 매우 크다고 할 수 있다. 이러한 사전 정보를 고려하여 판별 규칙을 정하는 임의의 고객이 우량 고객 집단 P_1 에 속할 사전 확률을 π_1 이라고 하고 불량 고객 P_2 에 속할 사전 확률이 π_2 라고 할 수 있는 것이다.

$$\pi_1 = P(\text{고객} \in P_1), \quad \pi_2 = P(\text{고객} \in P_2), \quad \pi_1 + \pi_2 = 1$$

사전확률을 고려하는 경우, 베이즈 판별 규칙은 다음과 같이 정의된다. 즉, 새로운 개체의 특성 $\boldsymbol{X} = \boldsymbol{x}$ 가 주어졌을 때, 각 집단에서 이 특성이 관측될 가능도(likelihood function)에 사전 확률을 곱한 값을 계산하여 더 큰 집단에 개체를 배정하는 방법이다. 이 경우 베이즈 판별 규칙(Bayes Discrimination Rule)은 다음과 같이 정의된다.

$$\boldsymbol{x} \in R_1 \text{ if } \pi_1 f_1(\boldsymbol{x}) > \pi_2 f_2(\boldsymbol{x}), \quad \text{and} \quad \boldsymbol{x} \in R_2 \text{ if } \pi_1 f_1(\boldsymbol{x}) \leq \pi_2 f_2(\boldsymbol{x})$$

위의 규칙에서 가능도 함수와 사전 확률의 곱을 사후 확률(posterior probability)이라고 한다. 사후 확률은 관측된 표본의 값이 주어진 경우 개체가 집단에 속할 확률을 의미한다.

$$P(P_i|\boldsymbol{X}=\boldsymbol{x}) \propto \pi_i f_i(\boldsymbol{x}), \quad i=1,2$$

5.1.4. 최적 판별 규칙

이제 판별에서 발생하는 오류에 대한 비용도 함께 고려할 수 있는 최적의 규칙을 고려해 보자. 앞 절에서 정의한 오류에 대한 비용 c_{12} 와 c_{21} 을 고려할 때 최적 판별 규칙은 다음에서 정의한 기대 오류비용 (Expected cost of misclassification; ECM) 을 최소로 하는 규칙이다.

$$\text{ECM} = c_{12}p_{12}\pi_2 + c_{21}p_{21}\pi_1 \tag{5.9}$$

기대 오류비용(ECM)을 최소로 하는 판별 규칙을 유도하면 다음과 같은 최적 판별 규칙을 얻을 수 있다.

$$R_1 = \left\{\boldsymbol{x}|\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \left[\frac{c_{12}}{c_{21}}\right]\left[\frac{\pi_2}{\pi_1}\right]\right\}, \quad R_2 = \left\{\boldsymbol{x}|\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \leq \left[\frac{c_{12}}{c_{21}}\right]\left[\frac{\pi_2}{\pi_1}\right]\right\} \tag{5.10}$$

식 5.10 에서 알 수 있듯이, 최대가능도 규칙과 베이즈 판별 규칙은 기대 오류비용을 최소화하는 최적 규칙의 특별한 경우이다. 즉, 비용이 동일한 경우 $c_{12} = c_{21}$ 에는 베이즈 판별 규칙과 동일하다. 또한 사전 확률도 동일한 경우 $\pi_1 = \pi_2$ 에는 최대 가능도 판별 규칙과 동일하다.

Exercise 5.1 (두 정규 분포의 판별 규칙)**.** 두 집단 P_1 과 P_2 에 속한 개체들의 특성 $\boldsymbol{X}$ 가 각각 다음과 같은 분산이 동일한 일변량 정규 분포를 따른다고 하자.

$$X|P_1 \sim N(\mu_1, \sigma^2), \quad X|P_2 \sim N(\mu_2, \sigma^2), \quad \mu_1 < \mu_2$$

이제 두 정규분포 확률밀도함수의 비, 최대가능도 함수의 비율을 고려해 보자.

$$\frac{f_1(x)}{f_2(x)} = \mathsf{exp}\left\{-\left(\frac{(x-\mu_1)^2}{2\sigma^2}\right) - \left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)\right\}$$

위의 식에서 두 정규분포의 비가 1 보다 큰 표본의 영역 R_1 을 구하면 다음과 같다.

$$\begin{aligned}
&\frac{f_1(x)}{f_2(x)} > 1 \\
&\iff -\frac{1}{2\sigma^2}(x-\mu_1)^2 + \frac{1}{2\sigma^2}(x-\mu_2)^2 > \mathsf{log}(1) \\
&\iff 2x\mu_1 - \mu_1^2 - 2x\mu_2 + \mu_2^2 > 0 \\
&\iff 2x(\mu_1-\mu_2) > \mu_1^2 - \mu_2^2 \\
&\iff x < \frac{\mu_1+\mu_2}{2}
\end{aligned}$$

따라서 두 집단의 평균의 중간에 있는 값 $(\mu_1+\mu_2)/2$ 보다 관측한 값 x 가 작으면 그룹 P_1 에 속한다고 결정한다.

$$R_1 = \left\{x|x < \frac{\mu_1 + \mu_2}{2}\right\}, \quad R_2 = \left\{x|x \geq \frac{\mu_1 + \mu_2}{2}\right\}$$

□

5.2. 다변량 정규분포와 판별 규칙

앞 절에서는 개체의 특성을 나타내는 확률변수가 일변량 정규분포인 경우의 예제를 보았다. 일반적인 경우 개체의 특성을 나타내는 특성값은 두 개 이상인 경우가 흔하다. 따라서 이제는 개체의 특성이 여러 개의 확률변수, 즉 확률벡터로 구성되어 있다고 가정하자.

p차원 확률벡터 $\boldsymbol{X}$ 를 고려하며 분포는 다변량 정규분포를 가정한다. 정규분포의 가정은 강한 가정이지만 단순하고 다루기 쉬운 분포이며 실제로 사용하기 용이하다. (하지만 조심해서 가정을 검토해야함)

두 개의 다변량정규분포를 따르는 두 집단 P_1, P_2으로 나누어져 있다고 가정하고 두 분포를 고려하자. 처음에는 문제를 쉽게 하기 위하여 평균은 다르고 공분산은 같다고 가정하자

$$P_1 : \ N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad P_2 : \ N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \quad \text{where } \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

참고로 다변량 정규 분포의 확률 밀도 함수는 다음과 같다.

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

식 5.9 에서 기대 오류비용 ECM 을 최소하는 판별규칙 식 5.10 과 같이 구할 수 있으며 이제 다변량 정규분포의 확률밀도함수를 식 5.10 에 넣고 판별함수를 유도해 보자.

$$\begin{aligned}
\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\right] \\
&= \exp\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2)\right] \\
&= \exp\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]
\end{aligned}$$

이제 ECM을 최소화하는 규칙은 개체의 특성이 $\boldsymbol{x} = \boldsymbol{x}_0$ 로 주어진 경우 다음을 만족하면 P_1으로 분류한다.

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) > \log\left(\frac{c_{12}}{c_{21}}\frac{\pi_2}{\pi_1}\right)$$

실제로 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$은 알 수 없기 때문에 표본 자료를 이용하여 추정을 한다.

$$\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{x}}_1 \quad \text{and} \quad \hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{x}}_2$$

$$\hat{\Sigma} = \boldsymbol{S}_p = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}$$

여기서 n_1, n_2는 표본에서 두 그룹에 대한 표본의 크기이며 $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$는 각각의 표본평균 벡터이다. $\boldsymbol{S}_p$는 두 개의 그룹에서 구한 공분산 행렬의 추정치를 결합한 합동 공분산 추정량이다.

위에서 구한 모수들의 추정치를 모집단으로 부터 구한 판별함수에 넣으면 다음과 같은 표본을 이용한 판별함수가 얻어진다.

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \boldsymbol{S}_p^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) > \mathsf{log}\left(\frac{c(1|2)}{c(2|1)}\frac{\pi_2}{\pi_1}\right) \tag{5.11}$$

위 식 5.11 의 표본 판별함수에서 아래와 같이 벡터 $\boldsymbol{a}$와 상수 m을 정의하면

$$\begin{aligned} \boldsymbol{a} &= \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\ m &= \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\boldsymbol{a}^t \bar{\boldsymbol{x}}_1 + \boldsymbol{a}^t \bar{\boldsymbol{x}}_2) \end{aligned} \tag{5.12}$$

ECM을 최소화하는 판별 규칙은 다음과 같이 특성값 $\boldsymbol{x}_0$의 **선형함수**로 나타낼 수 있다.

$$\boldsymbol{a}^t \boldsymbol{x}_0 > m + \mathsf{log}\left(\frac{c(1|2)}{c(2|1)}\frac{\pi_2}{\pi_1}\right) \tag{5.13}$$

만약 비용이 같고 ($c_{12} = c_{21}$) 두 집단에 대한 사전확률이 같다면 ($\pi_1 = \pi_2$) 판별함수는 다음과 같이 간단하게 나타낼 수 있다.

$$\boldsymbol{a}^t \boldsymbol{x}_0 > m \tag{5.14}$$

마지막으로 두 모집단의 공분산이 다를 경우는 판별함수가 확률벡터 $\boldsymbol{x}$의 선형함수로 나타나지 않는다.

5.3. Fisher의 선형 판별함수

Fisher(1938)는 위에서 다룬 ECM을 최소화하는 판별함수를 유도하는 방법과 완전히 다른 규칙을 사용하여 ECM을 최소화하는 방법과 동일한 결과를 유도하였다.

Fisher는 다변량 확률벡터 $\boldsymbol{x}$를 일변량 변수 y 로 선형변환하여 차원을 축소하는 방법을 고려하였다. 차원을 축소할 때 축소된 차원의 특성값 y가 두 집단을 최대로 구별되게 하는 선형변환을 구하는 문제를 생각하였다. Fisher의 방법은 특성값 벡터 $\boldsymbol{x}$에 확률적 분포가정을 고려하지 않는다.

특성값 벡터 $\boldsymbol{x}_1$을 집단 P_1의 특성값이라고 하고 $\boldsymbol{x}_2$을 집단 P_2의 특성벡터라고 하자. 이제 특성값 벡터 $\boldsymbol{x}_1$과 $\boldsymbol{x}_2$에 동시에 적용하는 선형변환 벡터를 $\boldsymbol{a}$라고 하고 각 집단의 일변량 특성값 y_1과 $ y_2$를 다음과 같이 정의하자.

$$y_1 = \boldsymbol{a}^t \boldsymbol{x}_1 \quad \text{and} \quad y_2 = \boldsymbol{a}^t \boldsymbol{x}_2$$

이제 두 집단에서 각각 n_1, n_2 개의 표본을 독립적으로 추출하였다거 가정하자.

$$y_{11}, y_{12}, \cdots, y_{1n_1} \sim_{ind} P_1 \quad \text{and} \quad y_{21}, y_{22}, \cdots, y_{2n_2} \sim_{ind} P_2$$

각 집단의 평균벡터는 다음과 같이 일변량 평균으로 변환된다.

$$\bar{y}_1 = \boldsymbol{a}^t \bar{\boldsymbol{x}}_1 \quad \text{and} \quad \bar{y}_2 = \boldsymbol{a}^t \bar{\boldsymbol{x}}_2$$

Fisher의 방법은 변환된 일변량 변수 y가 두 집단 간의 차이를 최대로, 집단 내의 차이를 최소로 하도록 하는 선형변환 $\boldsymbol{a}$를 유도하는 것이다.

두 집단 간의 차이를 최대로 하는 것은 두 평균의 차이 $|\bar{y}_1 - \bar{y}_2|$가 커지게 하는 것이고 집단 내의 차이를 최소로 하도록 하는 것은 집단내의 변동, 즉 집단내의 분산을 작게하는 것이다. 이러한 두 집단 간의 차이와 집단 내의 변동을 같이 반영할 수 있는 측도를 다음과 같이 생각하였다.

$$\max_{\boldsymbol{a}} \frac{|\bar{y}_1 - \bar{y}_2|}{s_y} = \max \frac{\text{variation between groups}}{\text{variation within group}} \tag{5.15}$$

여기서 $s^2_y $는 두 집단의 특성값 y의 합동분산 추정량이다.

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

이제 두 집단 간의 차이를 최대로, 집단 내의 변동를 최소로 하는 변환을 유도해보자. 식 5.15 에 제시된 값의 제곱이 가질 수 있는 상한(upper bound)을 유도하고 그 상한값을 취하는 선형변환 벡터 $\boldsymbol{a}$를 찾아보자.

여기서 두 집단의 평균벡터의 차이 $\boldsymbol{d}$를 다음과 같이 정의한다.

$$\boldsymbol{d} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

이제 식 5.15 에 주어진 값의 제곱에 대한 상한을 유도해보자.

$$\begin{aligned}
\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} &= \frac{(\boldsymbol{a}^t \bar{\boldsymbol{x}}_1 - \boldsymbol{a}^t \bar{\boldsymbol{x}}_2)^2}{\boldsymbol{a}^t \boldsymbol{S}_p \boldsymbol{a}} \\
&= \frac{[\boldsymbol{a}^t (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^2}{\boldsymbol{a}^t \boldsymbol{S}_p \boldsymbol{a}} \\
&= \frac{(\boldsymbol{a}^t \boldsymbol{d})^2}{\boldsymbol{a}^t \boldsymbol{S}_p \boldsymbol{a}} \\
&= \frac{(\boldsymbol{a}^t \boldsymbol{S}_p^{1/2} \boldsymbol{S}_p^{-1/2} \boldsymbol{d})^2}{\boldsymbol{a}^t \boldsymbol{S}_p \boldsymbol{a}} \\
&\leq \frac{(\boldsymbol{a}^t \boldsymbol{S}_p \boldsymbol{a})(\boldsymbol{d}^t \boldsymbol{S}_p^{-1} \boldsymbol{d})}{\boldsymbol{a}^t \boldsymbol{S}_p \boldsymbol{a}} \\
&= \boldsymbol{d}^t \boldsymbol{S}_p^{-1} \boldsymbol{d}
\end{aligned} \tag{5.16}$$

위에서 부등식은 코쉬-쉬바르쯔 부등식(Cauchy–Schwarz Inequality) 을 적용한 결과이며 부등식의 등호는 다음과 같은 경우에 성립한다.

$$\boldsymbol{a} = \boldsymbol{S}_p^{-1}\boldsymbol{d} = \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \tag{5.17}$$

코쉬-쉬바르쯔 부등식

두 벡터 $\boldsymbol{u}$와 $\boldsymbol{v}$에 대하여 다음 부등식(Cauchy–Schwarz Inequality)이 성립한다.

$$|\boldsymbol{u}^t\boldsymbol{v}|^2 \le (\boldsymbol{u}^t\boldsymbol{u})(\boldsymbol{v}^t\boldsymbol{v}) \tag{5.18}$$

부등식의 등호는 $\boldsymbol{u}$와 $\boldsymbol{v}$가 선형종속일 때, 즉 $\boldsymbol{u} = c\boldsymbol{v}$ 인 경우 성립한다.
식 5.16 에서 다음과 같이 벡터 $\boldsymbol{u}$ 와 $\boldsymbol{v}$ 를 적용하면 원하는 결과를 얻는다.

$$\boldsymbol{u} = \boldsymbol{S}_p^{1/2}\boldsymbol{a} \quad \text{and} \quad \boldsymbol{v} = \boldsymbol{S}_p^{-1/2}\boldsymbol{d}$$

식 5.17 에서 유도한 선형변환 벡터 $\boldsymbol{a}$는 다변량 정규분포 가정하에서 ECM을 최소화하는 판별함수의 선형벡터 식 5.12 와 동일하다.

더 나아가 Exercise 5.1 의 결과를 이용하면, 두 집단의 평균이 각각 $\bar{y}_1$ 과 $\bar{y}_2$ 이므로 새로운 관측값 $y_0 = \boldsymbol{a}^t\boldsymbol{x}_0$ 의 값을 두 평균의 가운데 값, 즉 $m = (\bar{y}_1 + \bar{y}_2)/2$와 비교하여 판별할 수 있다. 즉, 개체의 특성값 $y_0 = \boldsymbol{a}^t\boldsymbol{x}_0$일 때 다음을 만족하면 집단 P_1으로 분류한다.

$$y_0 > m$$

여기서

$$y_0 = \boldsymbol{a}^t\boldsymbol{x}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t\boldsymbol{S}_p^{-1}\boldsymbol{x}_0$$

$$m = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^t\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}\boldsymbol{a}^t(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

따라서 Fisher의 판별함수에 의한 분류 규칙은 다변량 정규분포 가정하에 ECM을 최소로 하는 규칙 식 5.14 과 동일하다.

5.4. 예제: 잔디깎는 트렉터

다음은 미국에서 한 소비용품 판매점이 잔디깎는 트랙터(lawn mower)의 소유 여부에 대한 가구의 소득과 집 크기의 관계를 알아보기 위하여 고객의 정보를 수집한 자료이다. 총 24개의 자료가 있고 트랙터를 소유하지 않는 집과 소유한 집이 각각 12개이다.

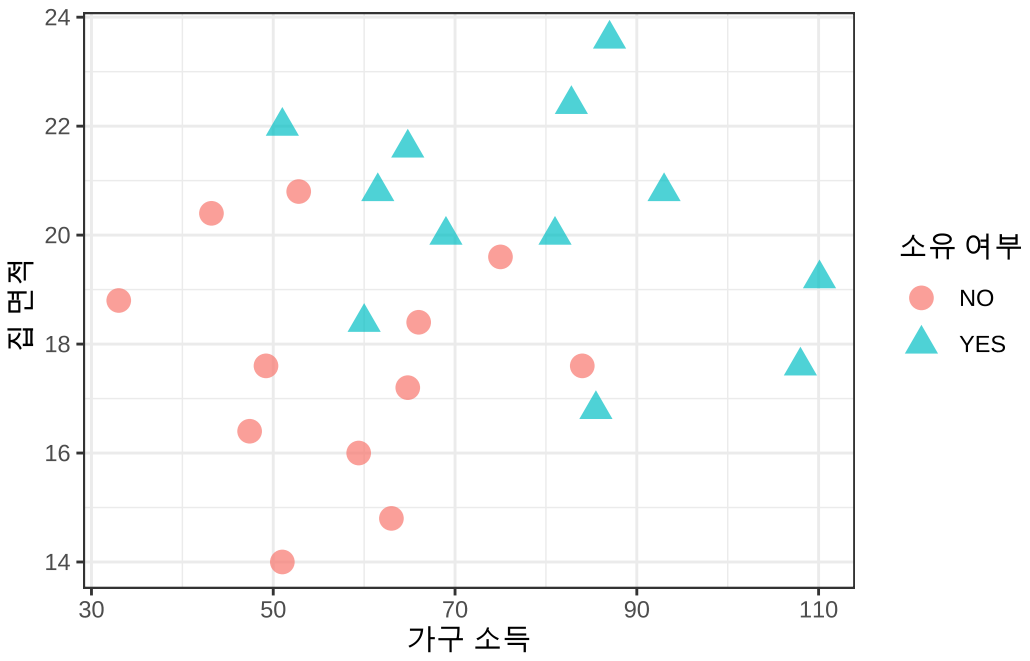
변수 income 과 lotsize 는 각각 가구 소득과 집의 크기에 대한 변수이다. 트랙터를 소유한 상태를 나타내는 변수는 class 이며 값이 2 이면 소유하지 않는 상태(NO,**그룹 1**), 1 이면 소유한 상태(YES, **그룹 2**) 를 나타낸다.

```
# read data
df <- read.csv(here("data","lawn.csv") , header=T, sep="")
df$class <- factor(df$class, levels=c(2,1), labels=c( "NO","YES"))
head(df)
```

```
  income lotsize class
1   60.0    18.4   YES
2   85.5    16.8   YES
3   64.8    21.6   YES
4   61.5    20.8   YES
5   87.0    23.6   YES
6  110.1    19.2   YES
```

이제 트렉터 소유의 여부와 소득/집크기의 관계를 알아보기 위하여 이차원에 자료를 그림으로 나타내보자.

```
# plot data wiyth different symbol by uisng ggplot2

df %>% ggplot(aes(x=income,y=lotsize, shape=class, color=class)) +
  geom_point(size=4, alpha = 0.7) +
  theme_bw() +
  labs(x="가구 소득",y="집 면적 ",shape="소유 여부", color = "소유 여부")
```



이제 앞절에서 유도한 선형판별함수를 트렉터 자료에 적용하여 판별함수를 구해보자. 일단 표본 평균벡터와 합동 공분산 행렬을 계산한다.

```
# 각 그룹의 표본 크기
n1 <- sum(df$class == "NO" )
n2 <- sum(df$class == "YES" )

n1
```

```
[1] 12
```

```
n2
```

```
[1] 12
```

```
# 각 그룹의 평균벡터
mean_vec <- df %>%
   group_by(class) %>%
   summarise(across(c(income, lotsize), mean))

mean_vec
```

```
# A tibble: 2 x 3
  class income lotsize
  <fct>  <dbl>   <dbl>
1 NO      57.4    17.6
2 YES     79.5    20.3
```

```
mean_x1 <- mean_vec %>% dplyr::filter(class == "NO" ) %>%
  dplyr::select(income, lotsize) %>%
  as.matrix() %>% t()

mean_x2 <- mean_vec %>% dplyr::filter(class == "YES" ) %>%
  dplyr::select(income, lotsize) %>%
  as.matrix() %>% t()

mean_x1
```

```
            [,1]
income   57.40000
lotsize  17.63333
```

```
mean_x2
```

```
            [,1]
income   79.47500
lotsize  20.26667
```

```
# 각 그룹의 공분산 행렬

cov_tbl <- df %>%
  group_by(class) %>%
  summarise(cov = list(cov(across(c(income, lotsize)))), .groups = "drop")

cov_x1 <- cov_tbl %>% filter(class == "NO" ) %>% pull(cov) %>% .[[1]]
cov_x1
```

```
             income   lotsize
income   200.705455 -2.589091
lotsize   -2.589091  4.464242
```

```
cov_x2 <- cov_tbl %>% filter(class == "YES" ) %>% pull(cov) %>% .[[1]]
cov_x2
```

```
            income    lotsize
income   352.64386 -11.818182
lotsize -11.81818   4.082424
```

```
# 합동 공분산 행렬
#cov_p <- ((n1-1)*cov_x1 + (n2-1)*cov_x2)/(n1+n2-2)
cov_p <- (n1*cov_x1 + n2*cov_x2)/(n1+n2)
cov_p
```

```
             income   lotsize
income   276.674659 -7.203636
lotsize   -7.203636  4.273333
```

이제 트렉터 자료에 적용하여 앞장에서 구한 최적의 판별 규칙을 유도해보자. 먼저 식 5.12 와 식 5.17 에 나타난 최적 변환 벡터 $\boldsymbol{a}$ 와 상수 m 는 다음과 같이 구할 수 있다.

```
# find discriminant function
a <- solve(cov_p) %*%  (mean_x1 - mean_x2)
a
```

```
              [,1]
income  -0.1002303
lotsize -0.7851847
```

```
m <- t(mean_x1 - mean_x2) %*% solve(cov_p) %*% (mean_x1 + mean_x2)/2
m
```

```
          [,1]
[1,] -21.73876
```

따라서 새로운 특성값 벡터를 $\boldsymbol{x}_0 = (x_{10},\ x_{20})^t$ 이라고 하면 ECM을 최소화하는 판별함수는 다음과 같이 주어진다. 즉, 다음 조건이 만족하면 그룹 1, 즉 트랙터를 사지 않는 그룹에 속한다.

$$(-0.1002303) \times (x_{10}) + (-0.7851847) \times (x_{20}) > -21.7387617$$

만약 `income` 과 `lotsize`가 각각 70, 16 이라면 다음과 같이 $\boldsymbol{a}^t\boldsymbol{x}_0$ 를 구할 수 있다.

```
x_0 <- matrix(c(70, 16),2,1)
x_0
```

```
     [,1]
[1,]   70
[2,]   16
```

```
t(a) %*% x_0
```

```
          [,1]
[1,] -19.57908
```

다시 쓰면

$$(-0.1002303) \times (70) + (-0.7851847) \times (16) = -19.5790766$$

따라서 -19.5790766 〉 -21.7387617 이므로 그룹 1 (`NO`) 에 속한다고 판별한다.즉 트랙터를 소유하지 않는 집단에 속한다고 판별한다.

```
t(a) %*% x_0 > m
```

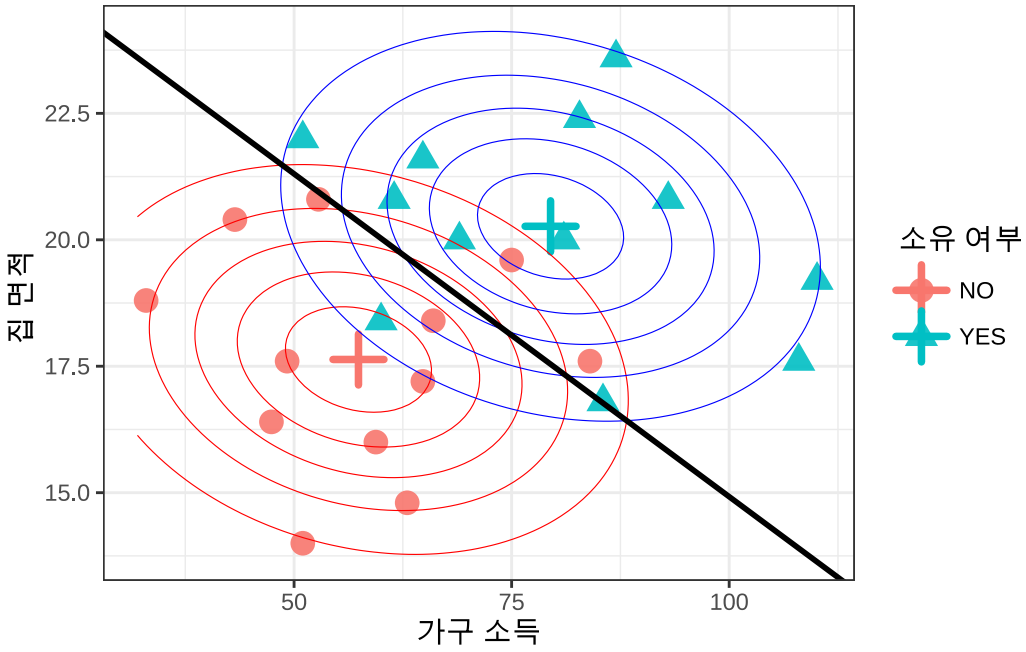
```
     [,1]
[1,] TRUE
```

즉, income 과 lotsize가 각각 70, 16 이면 $\boldsymbol{a}^t\boldsymbol{x}_0 > m$ 을 만족하므로 트렉터를 소유하지 않을 그룹 (NO) 에 속한다고 판별한다.

트랙터를 소유한 집단과 소유하지 않은 집단에서 소득과 집크기에 대하여 이변량 정규분포를 가정하고 각 집단의 평균벡터와 공통 공분산을 구하여 이변량 정규분포의 확률밀도함수를 이차원 평면에 표시해보았다. 또한 앞에서 구한 선형 판별함수의 경계선 $\boldsymbol{a}^t\boldsymbol{x} = m$ 을 표시하였다

```
# 등고선을 위한 격자점 생성
grid <- expand.grid(
  income = seq(min(df$income) - 1, max(df$income) + 1, length = 100),
  lotsize = seq(min(df$lotsize) - 1, max(df$lotsize) + 1, length = 100)
)
# 각 집단의 이변량 정규분포 확률밀도함수 계산
grid$dx1 <- dmvnorm(grid[,c("income","lotsize")], mean = mean_x1, sigma = cov_p)
grid$dx2 <- dmvnorm(grid[,c("income","lotsize")], mean = mean_x2, sigma = cov_p)

b0 <- as.numeric(m/a[2])
b1 <- -as.numeric(a[1])/as.numeric(a[2])

# 산점도와 등고선, 판별 경계선 그리기
df %>% ggplot(aes(x=income,y=lotsize, shape=class, color=class)) +
  geom_point(size=4, alpha = 0.9) +
  # 평균벡터 표기
  geom_point(data = mean_vec, aes(x=income, y=lotsize, color=class),  shape=3, size=5, stroke=
  # 등고선 1
  geom_contour(
    data = grid,
    aes(x = income, y = lotsize, z = dx1),
    color = "red", linewidth = 0.2, bins = 6,
    inherit.aes = FALSE
  ) +
  # 등고선 2
  geom_contour(
    data = grid,
    aes(x = income, y = lotsize, z = dx2),
    color = "blue", linewidth = 0.2, bins = 6,
    inherit.aes = FALSE
  ) +
  # 판별 경계선
  geom_abline(
    intercept = b0, slope = b1,
    color = "black", linewidth = 1
  ) +
```

```
  theme_bw() +
  labs(x="가구 소득",y="집 면적 ",shape="소유 여부", color = "소유 여부")
```



패키지 MAASS 의 함수 lda()을 이용하면 앞에서 구한 Fisher 의 선형판별함수를 쉽게 구할 수 있다.

```
fisher_model <- lda(class ~ income + lotsize, data = df, method="moment")
fisher_model
```

```
Call:
lda(class ~ income + lotsize, data = df, method = "moment")

Prior probabilities of groups:
 NO YES
0.5 0.5

Group means:
    income  lotsize
NO  57.400 17.63333
YES 79.475 20.26667

Coefficients of linear discriminants:
              LD1
income  0.0484468
lotsize 0.3795228
```

```
# 판별 벡터
fisher_model$scaling
```

```
              LD1
income  0.0484468
lotsize 0.3795228
```

참고로 함수 `lda` 의 결과값인 `Coefficients of linear discriminants` 는 식 5.17 에서 구한 최적 변환 벡터 $\boldsymbol{a}$ 와 같은 방향의 벡터이지만 길이가 다른 점에 유의하자.

```
# lda 함수의 결과와 비교
a/a[1]
```

```
            [,1]
income  1.000000
lotsize 7.833806
```

```
fisher_model$scaling/fisher_model$scaling[1]
```

```
             LD1
income  1.000000
lotsize 7.833806
```

이제 위에서 고려한 값, 즉 `income` 과 `lotsize`가 각각 70, 16 이라면 다음과 같이 새로운 데이터프레임을 만들고 `predict` 함수를 사용하여 앞에서 구한 판별 결과와 동일한 결과를 얻을 수 있다.

```
new_data <- data.frame(income = 70, lotsize = 16)
predict(fisher_model, newdata = new_data)
```

```
$class
[1] NO
Levels: NO YES

$posterior
         NO       YES
1 0.8965703 0.1034297

$x
        LD1
1 -1.043894
```

6. 주성분 분석

```
library(tidyverse)
library(here)
library(knitr)
library(mvtnorm)
library(ggfortify)
library(HSAUR2)
library(pheatmap)

#아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
#font_add_google("Nanum Pen Script", "gl")
font_add_google(name = "Noto Sans KR", family = "noto")
showtext_auto()
```

다변량 데이터를 다룰 때 문제점 중 하나는 단순히 변수가 너무 많아 데이터에 대한 유익한 초기 평가를 성공적으로 수행하기 어렵다는 점이다. 변수들 사이의 복잡한 상관관계는 분석에 어려움을 더한다. 또한 변수가 너무 많으면 연구자가 데이터에 적용하고자 하는 다른 다변량 기법에도 문제를 일으킬 수 있다.빅데이터가 흔한 시데에 데이터는 넘쳐나지만, 그 속에서 진짜 중요한 정보를 찾아내는 일은 점점 더 어려워지고 있다.

주성분 분석(PCA, Principal Component Analysis)은 단순히 하나의 통계 기법을 넘어, 데이터를 이해하는 새로운 방식을 제시한다. 주성분 분석은 다변량 자료의 변수들 속에 숨어 있는 공통된 구조를 발견하고, 이를 소수의 새로운 변수로 변환하는과정이다. 수십 개의 변수를 그대로 바라보면 혼란스러울 뿐이다. PCA는 이 변수들을 새로운 좌표계로 옮겨, 가장 큰 변동을 담아내는 방향으로 새로운 변수를 찾아낸다. 첫 번째 주성분이 데이터의 전체 흐름을 포착하고, 두 번째 주성분이 남은 변동을 설명한다. 이렇게 서로 직교하는 변수들이 순차적으로 만들어지면서, 데이터가 가지고 있는 변동을 2-3 개의 새로운 변수로 단순하게 많은 부분을 설명할 수 있다.

주성분 분석의 목적은 분명하다. 첫째, 차원을 축소하여 핵심적인 소수의 변수들만 남기는 것이다. 복잡한 다변량을 2-3 개의 주성분으로 요약하면, 분석은 훨씬 단순해진다. 주성분 분석은 자료가 많은 수의 변수들(Variables)로 구성되어 있는 경우 적은 개수의 변수로 자료의 차원을 축소하는 것이 주요한 목적이다(Dimension Reduction). 또한 이렇게 차원을 축소하는 경우 원래 변수들이 가지고 있던 변동(Variation)을 차원이 축소된 경우에도 최대한 유지할 수 있도록 한다.

둘째로 주성분 분석은 데이터의 시각화를 더욱 쉽게 수행할 수 있다. 수십 차원의 세계를 2차원, 3차원으로 투영해 보면, 군집은 또렷해지고 이상치는 드러난다. 셋째, 주어진 변수들의 복잡한 상관관계를 검토하고 분석할 필요가 없어진다. 서로 얽힌 변수들을 직교하는 주성분으로 바꿔놓으면, 다중공선성이라는 문제는 자연스레 사라진다.

주성분 분석은 특정 학문에만 국한되지 않는다. 사회과학에서 복잡한 설문 문항을 요약하는 데, 생물학에서 유전자 발현 데이터의 패턴을 찾는 데, 경영학에서 고객 행동의 핵심 요인을 도출하는 데, 더 나아가 인공지능에서 이미지나 텍스트의 고차원 정보를 다루는 데까지, 그 쓰임새는 한계가 없다.

6.1. 이변량 확률벡터의 변환

6.1.1. 이변량 정규분포

먼저 주성분분석의 기본개념을 이해하기 위하여 다음과 같은 평균 $\boldsymbol{\mu}$인 이차원 확률벡터 $\boldsymbol{X} = (X_1, X_2)^t$ 가 아래와 같이 공분산을 가지는 정규분포를 따른다고 하자.

$$\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad E(\boldsymbol{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = V(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

확률벡터 $\boldsymbol{X}$의 상관계수행렬 $\boldsymbol{C}$ 는 다음과 같이 표시한다.

$$\boldsymbol{C} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

참고할 사항은 상관계수행렬은 표준화된 확률변수 $Z_i = (X_i - \mu_i)/\sqrt{\sigma_{ii}}$의 공분산 행렬이다. 이유는 다음과 같이 보일 수 있다. 일단 표준화된 확률벡터 $\boldsymbol{Z} = (Z_1, Z_2)^t$를 정의하자.

$$\boldsymbol{Z} = \begin{bmatrix} (X_1 - \mu_1)/\sqrt{\sigma_{11}} \\ (X_2 - \mu_2)/\sqrt{\sigma_{22}} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1/\sqrt{\sigma_{22}} \end{bmatrix} \begin{bmatrix} (X_1 - \mu_1) \\ (X_2 - \mu_2) \end{bmatrix} = \boldsymbol{D}(\boldsymbol{X} - \boldsymbol{\mu})$$

여기서 $\boldsymbol{D}$는 각 변수의 표준편차의 역수를 대각원소로 가지는 대각 행렬이다. 표준화된 확률벡터 $\boldsymbol{Z}$의 공분산은 다음과 같이 유도된다.

$$\begin{aligned} V(\boldsymbol{Z}) &= E([\boldsymbol{D}(\boldsymbol{X} - \boldsymbol{\mu})][\boldsymbol{D}(\boldsymbol{X} - \boldsymbol{\mu})]^t) \\ &= \boldsymbol{D}E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^t)\boldsymbol{D}^t \\ &= \boldsymbol{D}\Sigma\boldsymbol{D}^t \\ &= \begin{bmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1\sqrt{\sigma_{22}} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1\sqrt{\sigma_{22}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \end{aligned}$$

6.1.2. 주성분의 기준과 생성방법

일반적으로 주성분 분석에서는 확률벡터 $\boldsymbol{X}$의 평균이 $\boldsymbol{0}$이라고 가정한다. 주성분 분석은 원래 변수들의 위치(location, 즉 평균)에 영향을 받는 방법이 아니라 자료의 변동(variation, 즉 분산)을 최대로 유지하는 새로운 변수를 만드는 것이 목적이다. 따라서 평균이 $\boldsymbol{0}$이 아닌 확률벡터 $\boldsymbol{X}_*$도 그 평균을 뺀 벡터 $\boldsymbol{X}$로 변환하여 평균을 $\boldsymbol{0}$

으로 만들고 주성분분석을 적용한다.

$$\text{Let } \boldsymbol{X} = \boldsymbol{X}_* - E(\boldsymbol{X}_*) = \boldsymbol{X}_* - \boldsymbol{\mu}_*, \text{ then } E(\boldsymbol{X}) = 0, \;\; V(\boldsymbol{X}) = V(\boldsymbol{X}_*)$$

아래부터는 특별한 언급이 없으면 확률벡터 $\boldsymbol{X}$가 평균이 $\mathbf{0}$이라고 가정한다.

이제 평균이 0이고 공분산(상관계수행렬)이 다음과 같이 주어지는 이변량정규분포를 생각해보자.

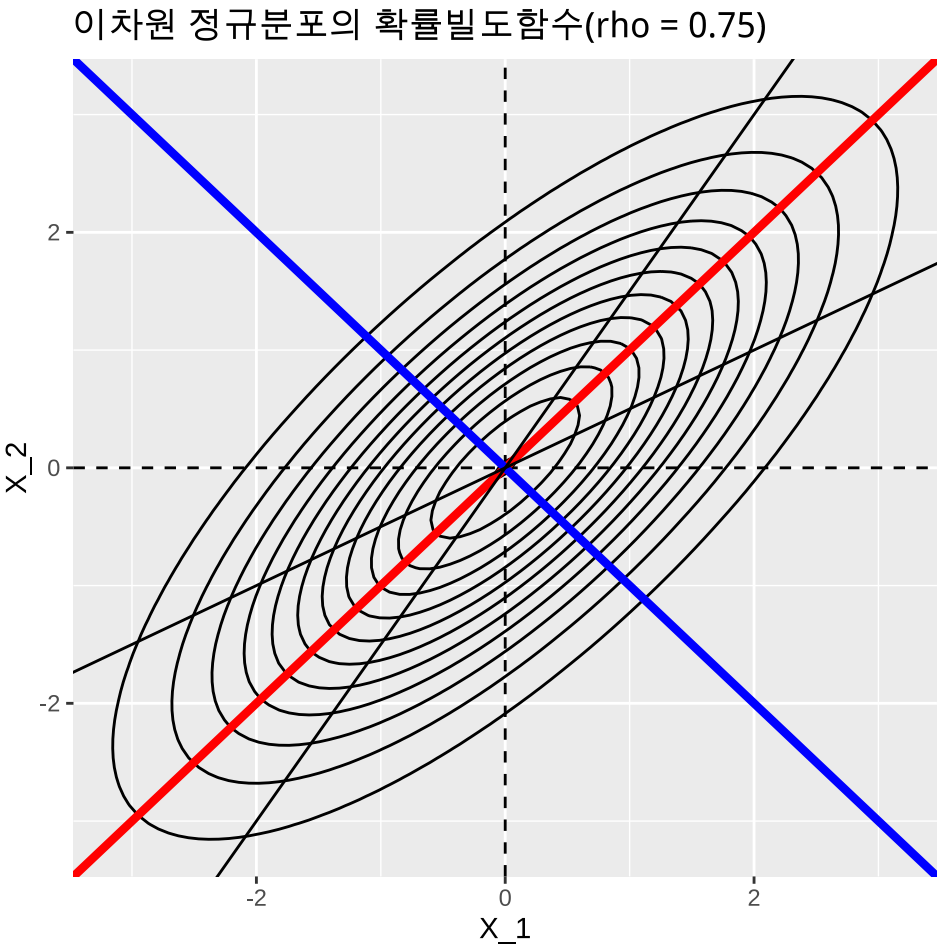
$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix} \tag{6.1}$$

참고로 두 변수의 상관계수는 $\rho = 1.5/2 = 0.75$ 이다

R 의 패키지 중 mtvnorm 을 사용하면 다변량분포에 대한 다양한작업을 손쉽게 할 수 있다. 다음 R 프로그램은 위에 주어진 공분산행렬 을 가지는 이변량정규분포 화률밀도함수의 2차원 등고선그림이다.

```
library(mvtnorm)
library(ggplot2)
library(dplyr)

#### 평균과 공분산
mu <- c(0, 0)
sigma <- matrix(c(2, 1.5, 1.5, 2), nrow = 2)

# grid 만들기
x.points <- seq(-4, 4, length.out = 100)
y.points <- seq(-4, 4, length.out = 100)

grid <- expand.grid(x1 = x.points, x2 = y.points)

#### 확률밀도 계산
grid$z <- dmvnorm(grid[, c("x1", "x2")], mean = mu, sigma = sigma)

#### ggplot으로 contour 그리기

p <- ggplot(grid, aes(x = x1, y = x2, z = z)) +
  geom_contour(color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_abline(slope = 1, intercept = 0, color = "red" ,linewidth = 1.5) +
  geom_abline(slope = -1, intercept = 0, color = "blue",linewidth = 1.5) +
  geom_abline(slope = 1.5, intercept = 0, color = "black") +
  geom_abline(slope = 0.5, intercept = 0, color = "black") +
  labs(x = "X_1", y = "X_2", title = "이차원 정규분포의 확률빌도함수(rho = 0.75) ")
```

```
print(p)
```



위의 그림에서 이차원 평면 상의 이차원 정규분포 밀도함수 등고선을 보면 분포의 퍼진 정도가 가장 큰 방향의 축이 원점을 지나고 기울기가 1 인 직선임을 알 수 있다(빨간 색 직선) 또한 이 직선과 직교하는 선은 원점을 지나고 기울기가 -1 인 직선임을 알 수 있다(파란 색 직선)

이제 각 확률변수 X_i 의 선형 변환으로 새로운 확률변수 Z_i 를 다음과 같이 정의한다.

$$Z_1 = a_{11}X_1 + a_{12}X_2, \quad Z_2 = a_{21}X_1 + a_{22}X_2$$

새로운 확률벡터 $\boldsymbol{Z} = (Z_1, Z_2)^t$ 는 다음과 같이 표시할 수 있다.

$$\boldsymbol{Z} = \boldsymbol{AX} \text{ where } \boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{6.2}$$

이렇게 새로운 확률 변수 Z_i를 만들 때 첫번째 변수 Z_1 를 첫번째 주성분(the 1st Principal Component)이라고 하며 원래 확률 벡터 $\boldsymbol{X}$가 가지는 총 변동 중에 최대한 큰 변동을 가질 수 있도록 만들고(빨간 색 직선 방향) 두번째 변수 Z_2 를 두번째 주성분(the 2nd Principal Component) 이라고 부르고 첫번째 변수 Z_1과 공분산이 0 이면서 나머지 변동을 가질 수 있도록(파란색 직선 방향) 만드는 방법이 주성분(principal components)을 만드는 기준이다.

1. $\max V(Z_1)$
2. $Cov(Z_1, Z_2) = 0$
3. $V(Z_1) + V(Z_2) = V(X_1) + V(X_2)$

위의 조건을 만족하는 선형변환 $\boldsymbol{Z}$는 공분산행렬 $\boldsymbol{\Sigma}$의 고유값과 고유벡터로 구할 수 있다. 식 6.1 에 주어진 공분산 행렬은 양정치 행렬이므로 고유값(eigen value) $\lambda_1 > \lambda_2 > 0$가 모두 양수이고 각 고유치에 대응하는 정규 직교 고유벡터(orthonormal eigen vector)의 행렬 $\boldsymbol{P}$을 이용하여 다음과 같은 분해가 가능하다.

$$\Sigma = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t \tag{6.3}$$

여기서 식 6.1 에 대해서 식 6.3 에 주어진 행렬의 분해를 구해보면 아래와 같다.

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 7/2 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad \boldsymbol{P} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

식 6.2 에서 정의된 선형변환을 다음과 같이 정의하면

$$\boldsymbol{A} = \boldsymbol{P}^t \text{ so that } \boldsymbol{Z} = \boldsymbol{P}^t\boldsymbol{X}$$

주성분은 다음과 같이 만들어 진다.

$$Z_1 = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2, \quad Z_2 = -\frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2$$

이렇게 만들어진 주성분 벡터의 공분산 행렬을 구해보자

$$\begin{aligned} V(\boldsymbol{Z}) &= V(\boldsymbol{P}^t\boldsymbol{X}) \\ &= \boldsymbol{P}^t V(\boldsymbol{X})\boldsymbol{P} \\ &= \boldsymbol{P}^t\boldsymbol{\Sigma}\boldsymbol{P} \\ &= \boldsymbol{P}^t\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t\boldsymbol{P} \\ &= \boldsymbol{\Lambda} \end{aligned}$$

위의 유도식에서 정규직교 고유벡터 $\boldsymbol{P}$는 직교행렬이므로 (i.e. $\boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$)

$$V(\boldsymbol{Z}) = \boldsymbol{\Lambda} = \begin{bmatrix} 7/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

따라서 첫번째 주성분의 분산은 가장 큰 고유치가 되고 두 번째 주성분의 분산은 두 번째 고유치가 되며 두 주성분의 공분산은 0이고 정규분포이므로 독립이다.

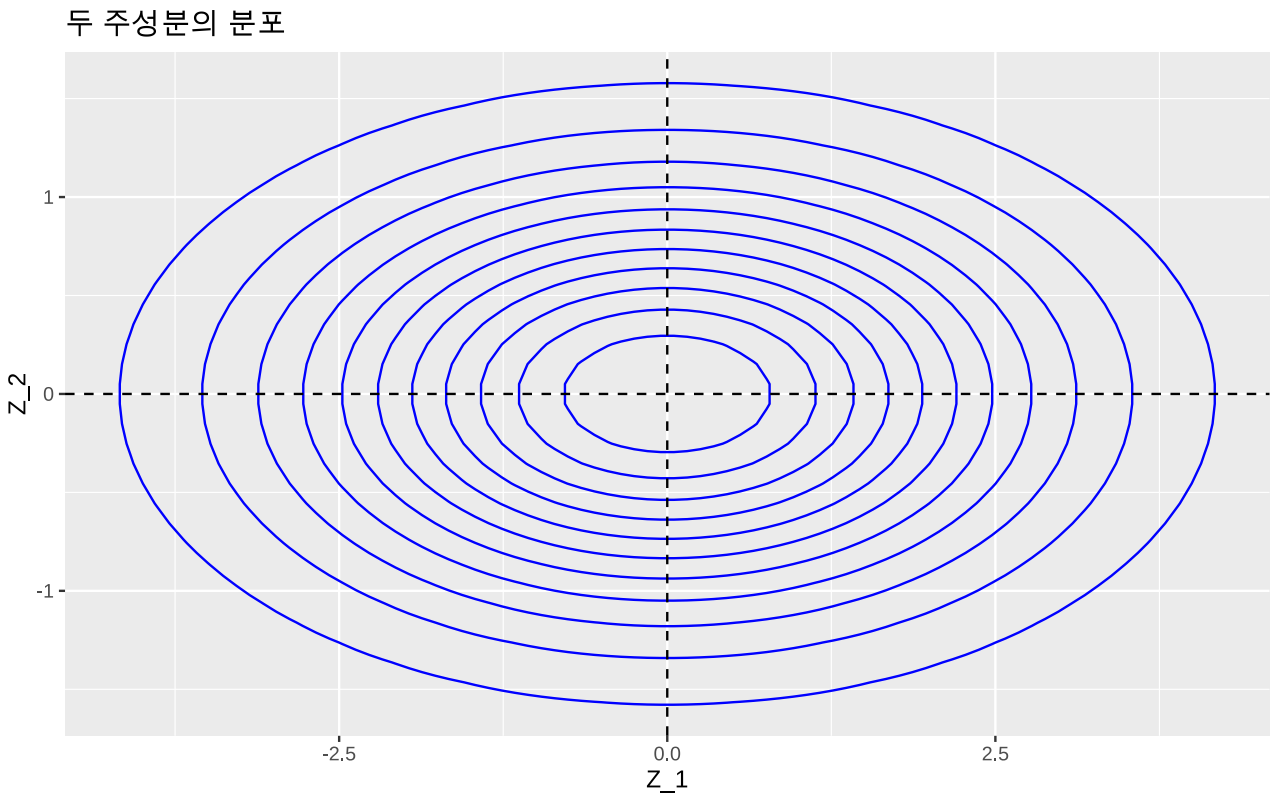
$$\boldsymbol{Z} \sim N_2(\boldsymbol{0}, \boldsymbol{\Lambda})$$

또한 다음의 관계가 성립한다.

$$V(Z_1) = 3.5, \;\; V(Z_2) = 0.5, \;\; Cov(Z_1, Z_2) = 0, \;\; V(Z_1) + V(Z_2) = V(X_1) + V(X_2)$$

다음은 주성분 벡터 $\boldsymbol{Z}$의 이변량정규분포의 2차원 등고선그림이다.

```
# 평균 벡터와 공분산 행렬
mu <- c(0, 0)
sigma1 <- matrix(c(3.5, 0,
                   0, 0.5), nrow = 2)

# x, y 격자 만들기
x.points <- seq(-5, 5, length.out = 100)
y.points <- seq(-5, 5, length.out = 100)
grid <- expand.grid(x = x.points, y = y.points)

# 이변량 정규분포 밀도 계산
grid$z <- dmvnorm(cbind(grid$x, grid$y), mean = mu, sigma = sigma1)

# ggplot2 시각화
p <- ggplot(grid, aes(x = x, y = y, z = z)) +
  geom_contour(color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  labs(x = "Z_1", y = "Z_2", title = "두 주성분의 분포")

print(p)
```



위와 같이 만든 첫번째 주성분은 원래 변수가 가지고 있는 총변동 $V(X_1)+V(X_2)=4$의 85.7%를 설명한다.

$$\frac{V(Z_1)}{V(X_1)+V(X_2)}=\frac{V(Z_1)}{V(Z_1)+V(Z_2)}=\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{3.5}{4}=0.857$$

원래의 두 변수 X_1과 X_2를 모두 사용하지 않고 하나의 주성분 Z_1만으로서 분포의 전체 변동의 큰 부분(85.7%)을 설명할 수 있다. 이러한 논리가 주성분을 이용한 차원의 축소이다.

6.1.3. 상관계수행렬을 통한 주성분분석

각 변수는 숫자로 나타나므로 측정할 경우 그 단위(unit)가 있다. 하지만 측정 단위들은 변수에 따라 또는 측정하는 사람에 따라 다를 수 있다. 예를 들어 키를 측정하는 경우 센티미터(cm)를 사용하고 몸무게를 측정하는 경우 킬로그램(Kg)를 사용한다. 또한 같은 변수인 키를 측정하는 경우에 센티미터(cm)대신 미터(m)를 사용할 수도 있다.

공분산행렬은 각 변수의 측정단위에 따라 변하며 그에 따른 고유값과 고유벡터도 변한다. 즉 주성분분석은 변수의 측정 단위에 따라 변할 수 있다.

예를 들어 식 6.1 을 공분산으로 가지는 확률벡터에서 첫번째 확률변수 X_1를 키라고 하고 그 측정단위가 미터(m)라고 가정하자. 만약 키의 측정단위를 센티미터로 바꾼다면 $(100\times X_1)$ 단위가 바뀐 확률벡터의 공분산행렬은 다음과 같이 변한다.

$$\mathbf{\Sigma}=\begin{bmatrix}(10000)(2) & (100)(1.5)\\ (100)(1.5) & 2\end{bmatrix}=\begin{bmatrix}20000 & 150\\ 150 & 2\end{bmatrix} \tag{6.4}$$

위에서 변환된 공분산행렬의 고유값과 고유벡터를 구해보면 다음과 같다.

```
sigma <- matrix(c(20000,150,150,2),nrow=2) # 공분산행렬
sigma
```

```
      [,1] [,2]
[1,] 20000  150
[2,]   150    2
```

```
eigen(sigma)
```

```
eigen() decomposition
$values
[1] 2.000113e+04 8.749508e-01

$vectors
              [,1]         [,2]
[1,] -0.999971874  0.007500117
[2,] -0.007500117 -0.999971874
```

가장 큰 고유치는 $\lambda_1 = 20000$에 가깝고 다음 고유치는 거의 0이다 ($\lambda_2 \sim 0$) 또한 고유벡터는 거의 단위행렬에 가까우므로 첫번째 주성분은 키를 나타내는 원래 변수와 동일하고(부호만 바뀐다) 전체 변동을 대부분 설명하며 두번째 주성분은 거의 설명하는 변동이 없게된다.

위와 같은 현상은 공분산행렬을 주성분분석에 사용하는 경우에 나타나는 큰 문제점이다. 많은 경우 변수들은 측정단위가 다르며 단위를 바꾸면 주성분이 크게 변한다. 이러한 분제점을 해결하기 위하여 많은 경우 각 변수를 표준화하여 사용하는 것이 좋다. 표준화하면 각 변수가 가지는 변동이 같게되며 서로의 관계는 상관계수로 파악할 수 있다. 앞에서 표준화된 변수들의 공분산행렬은 상관계수행렬임을 보였다. 따라서 특별한 이유가 없는 한 주성분분석은 상관계수 행렬을 사용한다.

만약 두개의 확률벡터가 각각 식 6.1 과 식 6.4 를 공분산으로 갖는다고 가정하자. 두 공분산행렬은 다르지만 상관계수행렬은 같음을 알수 있다.

$$\boldsymbol{C} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix} \tag{6.5}$$

따라서 단위의 종류와 변환에 관계없이 같은 주성분을 가진다.

```
sigma <- matrix(c(1,0.75,0.75,1),nrow=2) # 상관계수행렬(상관계수=0.75)
eigen(sigma)
```

```
eigen() decomposition
$values
[1] 1.75 0.25

$vectors
           [,1]       [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
```

위에서 구한 고유치와 고유벡터를 사용한 주성분과 그 통계적 성질은 다음과 같다.

$$Z_1 = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2, \quad Z_2 = -\frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2$$

$$V(Z_1) = 1.75, \;\; V(Z_2) = 0.25, \;\; Cov(Z_1, Z_2) = 0, \;\; V(Z_1) + V(Z_2) = V(X_1) + V(X_2) = 2$$

참고할 점은 상관계수행렬의 대각의 합은 변수의 개수($p = 2$)와 같고 상관계수가 변한다 하더라도 고유벡터들는 변하지 않아 주성분은 동일하게 정의되지만 각각의 분산은 달라진다. 아래는 상관계수 0.95를 가지는 상관계수행렬의 고유치와 고유값이다.

```
sigma <- matrix(c(1,0.95,0.95,1),nrow=2) # 상관계수행렬(상관계수=0.95)
sigma
```

```
     [,1] [,2]
[1,] 1.00 0.95
[2,] 0.95 1.00
```

```
eigen(sigma)
```

```
eigen() decomposition
$values
[1] 1.95 0.05

$vectors
           [,1]       [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
```

6.1.4. 표본자료를 이용한 주성분분석

이변량 정규분포에서 n개의 자료 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$이 독립표본으로 추출되었다면 표본공분산과 표본상관계수를 추정하여 주성분을 만들 수 있다.

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} \end{bmatrix} \quad \hat{\boldsymbol{C}} = \begin{bmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{bmatrix}$$

다음은 평균이 $\mathbf{0}$ 이고 다음을 공분산으로 가지는 이변량 정규분포에서 50개의 표본을 추출하고 표본상관계수를 추정하여 주성분을 만드는 R 프로그램이다.

$$\mathbf{\Sigma} = \begin{bmatrix} 3 & -2 \\ -2 & 2 \end{bmatrix}$$

```
mu <- c(0,0)  # 평균 벡터
sigma <- matrix(c(3,-2,-2,2),nrow=2) # 공분산행렬
X <- rmvnorm(100,mean=mu,sigma=sigma)
head(X)
```

```
           [,1]       [,2]
[1,]  4.6282705 -4.2243072
[2,]  1.4045731 -0.8958122
[3,]  0.4338715 -0.3901536
[4,]  0.9394071 -1.7978767
[5,] -2.1498722  1.7714280
[6,]  3.1633948 -1.8869771
```

```
C <- cor(X)
C
```

```
           [,1]       [,2]
[1,]  1.0000000 -0.8164816
[2,] -0.8164816  1.0000000
```

```
eigen(C)
```

```
eigen() decomposition
$values
[1] 1.8164816 0.1835184

$vectors
           [,1]       [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068
```

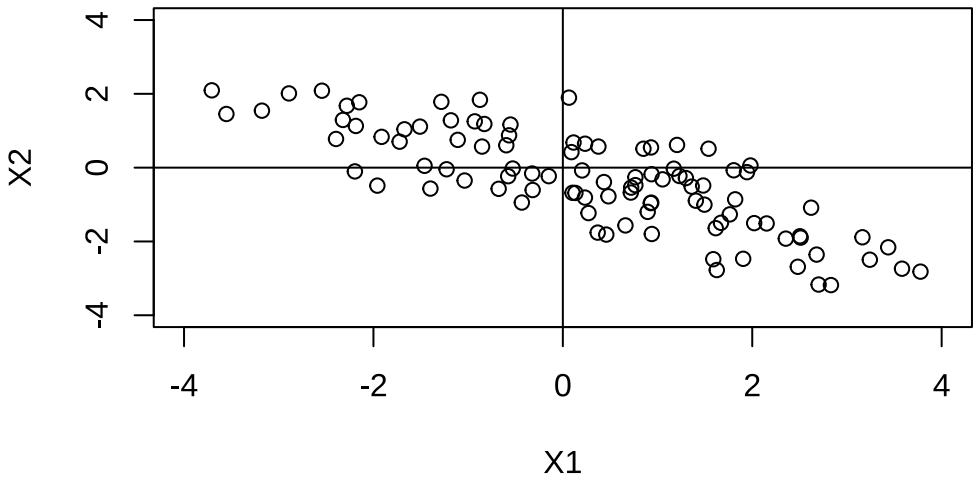
위의 결과로 주성분은 다음과 같이 주어진다.

$$Z_1 = -0.707X_1 + 0.707X_2 \quad Z_2 = -0.707X_1 - 0.707X_2$$

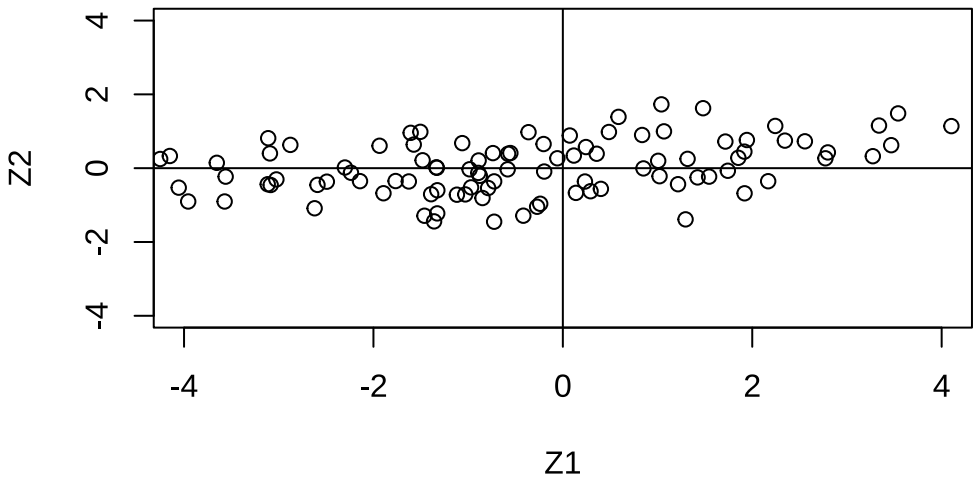
첫번째 주성분의 분산이 $V(Z_1) = \lambda_1 = 1.816$ 이므로 총변동 $p = 2$ 의 91%를 설명한다.

원래 자료 (X_1, X_2)의 산점도와 위의 식에 원래 자료를 넣어 계산된 주성분의 값 (Z_1, Z_2) 산점도는 다음 그림과 같다.

```
P <- eigen(C)$vectors
Z <- X %*% P
plot(X, xlab="X1",ylab="X2",ylim=c(-4,4),xlim=c(-4,4))
abline(v=0);abline(h=0)
```




```
plot(Z, xlab="Z1",ylab="Z2",ylim=c(-4,4),xlim=c(-4,4))
abline(v=0);abline(h=0)
```



6.2. 주성분 분석의 기초이론

주성분분석은 서로 상관된 p 개의 변수들 $\boldsymbol{X} = (X_1, X_2, \dots, X_p)^t$ 의 총 변동을 각 변수들의 선형결합으로 만든 서로 상관되지 않은 새로운 변수들인 $\boldsymbol{Z}^t = (Z_1, Z_2, \dots, Z_p)$에 의해 설명하는 것이다.

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + a_{i3}X_3 + \cdots + a_{ip}X_p, \quad i = 1, 2, \dots, p$$

주성분 분석의 목적은 원래의 관측변수들이 아닌 새로운 변수를 만들어서 원래 변수의 수보다 적은 수의 새로운 변수를 이용하여 원래 변수가 가지는 대부분의 변동을 설명하려는것이다. 즉, 차원의 축소(dimension reduction)가 그 목적이다.

새로운 변수인 Z_i들을 주성분(principal component, PC)이라고 부르며 X_i들의 모든 선형결합중에서 서로 상관되지 않고 원래 자료의 변동을 가능한 한 많이 설명한다는 의미에서 중요성이 감소하는 순서되로 유도된다.

즉 첫번째 주성분인 Z_1은 다른 주성분보다 분산이 가장 크다. 다음으로 두번째 주성분 Z_2는 Z_1과 상관관계가 없으며 남아있는 변동을 가능한 한 많이 설명하게 만든다. 이러한 작업을 계속 적용하여 p개의 주성분을 만들 수 있다.

$$V(Z_1) \geq V(Z_2) \geq \dots \geq V(Z_p) \quad \text{and} \quad Cov(Z_i, Z_j) = 0, \;\; i \neq j$$

주성분 분석은 다음의 경우에 매우 우용하다.

- 관측된 개수에 비하여 설명변수가 너무 많다.
- 설명변수들이 높게 상관되어 있다.

6.2.1. 주성분의 정의

$\boldsymbol{X} = (X_1, X_2, \dots, X_p)^t$를 차원이 p인 확률벡터(random vector)라고 하자.

$$E(\boldsymbol{X}) = \mathbf{0} \quad \text{and} \quad V(X) = \boldsymbol{\Sigma}$$

새로운 p개의 주성분 $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_p)^t$는 X_i들의 선형결합으로 만들어진다.

$$\begin{aligned}
Z_1 &= \boldsymbol{a}_1^t \boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1p}X_p \\
Z_2 &= \boldsymbol{a}_2^t \boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{2p}X_p \\
Z_3 &= \boldsymbol{a}_3^t \boldsymbol{X} = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + \cdots + a_{3p}X_p \\
&\dots \\
Z_p &= \boldsymbol{a}_p^t \boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + a_{p3}X_3 + \cdots + a_{pp}X_p
\end{aligned} \tag{6.6}$$

식 6.6 에서 정의된 주성분은 다음과 같은 통계적 성질을 가지고 있다.

$$E(Z_i) = \boldsymbol{a}_i^t \mathbf{0} = \mathbf{0} \quad \text{and} \quad V(Z_i) = \boldsymbol{a}_i^t \boldsymbol{\Sigma} \boldsymbol{a}_i \quad \text{and} \quad Cov(Z_i, Z_j) = \boldsymbol{a}_i^t \boldsymbol{\Sigma} \boldsymbol{a}_j$$

만약 주성분을 만드는 계수 벡터 $\boldsymbol{a}_i$들을 다음과 같이 행렬로 표시하면

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ & \dots & \dots & \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^t \\ \boldsymbol{a}_2^t \\ \dots \\ \boldsymbol{a}_1^t \end{bmatrix} \tag{6.7}$$

식 6.6 의 주성분 벡터 $\boldsymbol{Z}$는 다음과 같이 정의된다.

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X} \tag{6.8}$$

이제 주성분을 만드는 절차는 다음과 같다.

1.첫번째 주성분 $Z_1 = \boldsymbol{a}_1^t \boldsymbol{X}$ 를 다음과 같은 조건이 만족하게 찾는다.

- Maximize $Var(Z_1) = \boldsymbol{a}_1^t \boldsymbol{\Sigma} \boldsymbol{a}_1$
- subject to $\boldsymbol{a}_1^t \boldsymbol{a}_1 = 1$, i,e,

$$a_{11}^2 + a_{12}^2 + a_{13}^2 + \cdots + a_{1p}^2 = 1$$

2. 두번째 주성분 $Z_2 = \boldsymbol{a}_2^t \boldsymbol{X}$ 다음과 같은 조건이 만족하게 찾는다.

- Maximize $Var(Z_2) = \boldsymbol{a}_2^t \boldsymbol{\Sigma} \boldsymbol{a}_2$
- subject to $a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1$

- Z_2 is uncorrelated with Z_1, i.e.

$$Cov(Z_1, Z_2) = \boldsymbol{a}_1^t \boldsymbol{\Sigma} \boldsymbol{a}_2 = 0$$

....

k. k 번째 주성분 $Z_k = \boldsymbol{a}_k^t \boldsymbol{X}$ 다음과 같은 조건이 만족하게 찾는다.

- Maximize $Var(Z_k) = \boldsymbol{a}_k^t \boldsymbol{\Sigma} \boldsymbol{a}_k$
- subject to $a_{k1}^2 + a_{k2}^2 + \cdots + a_{kp}^2 = 1$
- Z_k is uncorrelated with $Z_1, Z_2, ..., Z_{k-1}$,

$$Cov(Z_k, Z_i) = \boldsymbol{a}_k^t \boldsymbol{\Sigma} \boldsymbol{a}_i = 0 \quad \text{for} \quad i = 1, 2, ..., k-1$$

주성분의 계수를 찾을 때 실제로는 공분산행렬 $\boldsymbol{\Sigma}$을 모르기 때문에 표본으로부터 얻어진 표본 공분산행렬 $\hat{\boldsymbol{\Sigma}} \equiv \boldsymbol{S}$를 이용한다. 또는 표본 상관계수행렬 $\hat{\boldsymbol{C}}$를 이용한다.

6.2.2. 양정치 행렬의 스펙트럼분해

주성분분석에서 선형결합에 이용되는 계수 a_{ij}는 공분산행렬의 스펙트럼 분해(Spectral Decomposition)을 통하여 찾을 수 있다 (식 B.4 참조).

공분산행렬은 대칭인 양정치행렬(positive definite matrix)이기 때문에 양의 실수를 가지는 고유값(eigenvalues)을 가진다. $p \times p$ 공분산 행렬을 $\boldsymbol{\Sigma}$라하면 다음과 같은 스펙트럼 분해가 가능하다.

$$\boldsymbol{\Sigma} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}^t$$

여기서 $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_q)$는 대각원소가 공분산행렬의 고유값인 대각행렬이며 고유값은 다음과 같이 내림차순으로 정렬되어 있다고 하자.

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$$

행렬 $\boldsymbol{P}$는 $p \times p$ 직교행렬(orthonomal matrix)로서 k번째 열벡터 $\boldsymbol{a}_k$는 k 번째 고유값에 해당하는 고유벡터이다.

$$\boldsymbol{\Sigma} \boldsymbol{a}_k = \lambda_k \boldsymbol{a}_k \quad \text{and} \quad a_{k1}^2 + a_{k2}^2 + \cdots + a_{kp}^2 = 1, \quad k = 1, 2, ..., p$$

$$\boldsymbol{P} = [\boldsymbol{a}_1 \ \ \boldsymbol{a}_2 \ \ ... \ \ \boldsymbol{a}_p]$$

여기서 행렬 $\boldsymbol{P}$는 직교행렬이므로 각 고유벡터들은 다음과 같은 관계가 있다.

$$\begin{cases} \boldsymbol{a}_i^t \boldsymbol{a}_j = 0 & \text{if} \quad i \neq j \\ \boldsymbol{a}_i^t \boldsymbol{a}_i = 1 & \text{if} \quad i = j \end{cases}$$

6.2.3. 이차형식의 최대값

주어진 대칭인 양정치행렬 $\boldsymbol{\Sigma}$에 대하여 다음과 같이 2차형식(quadratic form) $\boldsymbol{x}^t \Sigma \boldsymbol{x}$의 값을 최대로 하는 p-차원의 벡터 $\boldsymbol{x}$를 찾는 문제를 풀어보자.

$$\max_{\boldsymbol{x}} \frac{\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^t \boldsymbol{x}} = \max_{\boldsymbol{x}; |\boldsymbol{x}|=1} \boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x}$$

위의 문제에서 2차형식 $\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x}$는 벡터 $\boldsymbol{x}$의 길이를 증가시키면 비례하여 증가하므로 이차형식을 벡터의 길이의 제곱 $\boldsymbol{x}^t \boldsymbol{x}$로 나누어 표준화시킨 다음 그 양을 최대화하는 벡터를 찾는 문제로 생각하는 것이다.

2차형식(quadratic form) $\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x}$은 다음과 같은 스펙트럼분해가 가능하다.

$$\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x} = \boldsymbol{x}^t \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}^t x = \boldsymbol{x}^t [\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \dots \;\; \boldsymbol{a}_p] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} [\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \dots \;\; \boldsymbol{a}_p]^t \boldsymbol{x}$$

새로운 변수 $\boldsymbol{z} = \boldsymbol{P}^t \boldsymbol{x}$를 정의하면 2차형식은 다음과 같이 표시된다.

$$\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x} = \boldsymbol{z}^t \boldsymbol{\Lambda} \boldsymbol{z} = \lambda_1 z_1^2 + \lambda_2 z_2^2 + \dots \lambda_p z_p^2$$

고유값은 다음과 같은 관계를 가지므로

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

다음과 같은 부등식을 얻을 수 있으며

$$\lambda_p \sum_{i=1}^{p} z_i^2 \leq \sum_{i=1}^{p} \lambda_i z_i^2 \leq \lambda_1 \sum_{i=1}^{p} z_i^2$$

다시 위의 부등식에서 모든 항을 $\boldsymbol{z}^t \boldsymbol{z} = \sum_{i=1}^{p} \boldsymbol{z}_i^2$으로 나누면 다음의 부등식을 얻는다.

$$\lambda_p \leq \frac{\sum_{i=1}^{p} \lambda_i z_i^2}{\sum_{i=1}^{p} z_i^2} \leq \lambda_1$$

또한 $\boldsymbol{z}^t \boldsymbol{z} = \boldsymbol{x}^t \boldsymbol{P} \boldsymbol{P}^t \boldsymbol{x}^t = \boldsymbol{x}^t \boldsymbol{x}$으므로 마지막으로 아래의 부등식을 얻는다.

$$\lambda_p \leq \frac{\boldsymbol{x}^t \boldsymbol{\Sigma} \boldsymbol{x}}{\boldsymbol{x}^t \boldsymbol{x}} \leq \lambda_1$$

이 부등식의 의미는 표준화된 2차형식의 최대값은 가장 큰 고유값이며 최소값은 가장 작은 고유값이다.

6.2.4. 주성분의 계수

p-차원의 벡터 $\boldsymbol{a}$의 선형결합으로 새로운 변수 $Z_1 = \boldsymbol{a}^t\boldsymbol{X}$를 고려할때 Z_1의 분산은 다음과 같다.

$$Var(Z_1) = Var(\boldsymbol{a}^t\boldsymbol{X}) = \boldsymbol{a}^t Var(\boldsymbol{X})\boldsymbol{a} = \boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}$$

위에서 얻은 이차형식의 최대값을 구하는 과정을 보면 선형결합의 계수 $\boldsymbol{a}$를 공분산행렬에 대한 가장 큰 고유값 λ_1에 해당하는 표준고유벡터 $\boldsymbol{a}_1$를 사용하면 Z_1의 분산을 최대화 할 수 있다.

$$Var(Z_1) = Var(\boldsymbol{a}_1^t\boldsymbol{X}) = \boldsymbol{a}_1^t\boldsymbol{\Sigma}\boldsymbol{a}_1 = \boldsymbol{a}_1^t(\lambda_1\boldsymbol{a}_1) = \lambda_1$$

이렇게 첫 번째 주성분 $Z_1 = \boldsymbol{a}_1^t\boldsymbol{X}$을 찾을 수 있으며 두 번째 주성분, 세 번째 주성분들도 유사한 방법을 통해 찾아나간다. 두번째 주성분은 공분산행렬에 대한 두 번째로 큰 고유값 λ_2에 해당하는 표준 고유벡터 $\boldsymbol{a}_2$로 사용하면 그 분산은 λ_2이고 첫 번째 주성분과의 상관관계는 0이다.

$$Var(Z_2) = Var(\boldsymbol{a}_2^t\boldsymbol{X}) = \boldsymbol{a}_2^t Var(\boldsymbol{X})\boldsymbol{a}_2 = \lambda_2$$

$$Cov(Z_1, Z_2) = Cov(\boldsymbol{a}_1^t\boldsymbol{X}, \boldsymbol{a}_2^t\boldsymbol{X}) = \boldsymbol{a}_1^t Cov(\boldsymbol{X})\boldsymbol{a}_2 = \boldsymbol{a}_1^t\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^t\boldsymbol{a}_2 = 0$$

따라서 주성분분석에서 i번째 주성분의 계수는 공분산행렬 $\boldsymbol{\Sigma}$의 고유값 중 i 번째 큰 값인 λ_i에 해당하는 고유벡터 $\boldsymbol{a}_i$로 놓는다.

$$Z_i = \boldsymbol{a}_i^t\boldsymbol{X}, \quad i = 1, 2, \dots, p$$

6.2.5. 공분산행렬과 상관계수행렬

앞절에서 설명한 바와 같이 주성분을 반드는 경우 공분산행렬(Covariance matrix)를 사용하면 원래 변수의 단위가 바뀌면 주성분의 계수도 영향을 받는다. 변수들 중 분산이 큰 변수가 있으면 그 변수에 해당하는 주성분의 계수가 커지는 현상이 발생한다.

이러한 단점을 보완하기 위하여 보통의 경우 주성분은 공분산행렬이 아닌 상관계수행렬 $\boldsymbol{C}$ (Correlation matrix)의 고유값과 고유벡터를 이용하여 주성분을 만든다.

앞에서 전개한 주성분분석의 이론은 공분산행렬 $\boldsymbol{\Sigma}$를 상관계수행렬 $\boldsymbol{C}$ 로 바꾸어 전개하면 된다. 주의할 점은 상관계수행렬을 이용하는 경우 모든 대각원소가 1으므로 총변동은 변수의 개수 p가 된다.

$$\sum_{i=1}^{p} V(X_i) = \sum_{i=1}^{p} V(Z_i) = trace(\boldsymbol{C}) = p$$

6.3. 주성분 분석

6.3.1. 주성분 개수의 선택

주성분은 변수들의 모든 선형결합중에서 서로 상관되지 않고 원래 자료의 변동을 가능한 한 많이 설명한다는 의미에서 성분의 중요성이 감소하는 순서되로 유도된다. 따라서 주성분 분석의 목적에 따라 원래의 변수 개수인 p보다 적은 개수의 주성분들을 선택하여 원래 변수가 가지고 있는 변동의 많은 부분을 설명하려고 한다.

문제는 몇 개의 주성분을 선택해야 하는 문제이다.

각 주성분의 분산은 공분산 $\mathbf{\Sigma}$의 고유값이고 그 고유값들은 다음과 같은 순서로 정렬되어 있다

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$$

따라서 원래 변수 $X_1, X_2, ..., X_p$들의 분산의 합은 모든 고유값의 합과 같다.

$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Z_i)$$

이는 다음과 같은 식에서 유도 된다.

$$trace(\mathbf{\Sigma}) = trace(\boldsymbol{P}\mathbf{\Lambda}\boldsymbol{P}^t) = trace(\mathbf{\Lambda}\boldsymbol{P}^t\boldsymbol{P}) = trace(\mathbf{\Lambda})$$

따라서 주성분 Z_i는 자료의 총변동 $\sum_{i=1}^{p} \lambda_i = trace(\mathbf{\Sigma})$를 다음과 같은 비율만큼 설명한다.

$$P_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{6.9}$$

그리고 만약 $m < p$개의 주성분 $Z_1, Z_2, ..., Z_m$ 을 선택했다면 자료의 총변동을 다음과 같은 비율만큼 설명한다.

$$P_1 + P_2 + \cdots + P_m = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_m}{\sum_{j=1}^{p} \lambda_j} \tag{6.10}$$

주성분의 개수를 선택하는 방법은 매우 다양하며 다음과 같은 몇 개의 절차가 일번적인 방법이다.

- 총변동의 70%에서 90%를 설명할 수 있게 주성분의 개수를 선택

$$0.7 \leq P_1 + P_2 + \cdots + P_m \leq 0.9$$

- 평균고유값 $\sum_{i=1}^{p} \lambda_i / p$보다 작은 분산을 가지는 주성분을 제외한다.
- 주성분이 상관행렬로부터 추출되는 경우 고유값이 1보다 작은 주성분을 제외한다.

- Scree diagram을 이용: Scree diagram은 고유값을 큰 순서대로 그림으로 그리고 선을 연결항 그래프이다. 고유값의 감소가 느리게 진행되는 때까지 큰 고유치를 선택하면 된다.

6.3.2. 주성분의 척도변경

원래의 변수벡터 $\boldsymbol{X}$와 주성분 Z_j의 공분산은 다음과 같다.

$$Cov(\boldsymbol{X}, Z_j) = Cov(\boldsymbol{X}, \boldsymbol{a}_j^t \boldsymbol{X}) = Cov(\boldsymbol{X}, \boldsymbol{X}^t)\boldsymbol{a}_j = \boldsymbol{\Sigma a}_j$$

고유벡터 $\boldsymbol{a}_j$의 성질은 $ **a**_j = _j **a**_j$이므로 다음과 같은 결과를 얻고

$$Cov(\boldsymbol{X}, Z_j) = \boldsymbol{\Sigma a}_j = \lambda_j \boldsymbol{a}_j$$

따라서 원래의 변수 X_i와 주성분 Z_j의 공분산은 다음과 같다.

$$Cov(X_i, Z_j) = \lambda_j a_{ji}$$

더 나아가 원래의 변수 X_i와 주성분 Z_j의 상관계수는 다음과 같다.

$$Corr(X_i, Z_j) = \frac{Cov(X_i, Z_j)}{\sqrt{Var(X_i)Var(Z_j)}} = \frac{\lambda_j a_{ji}}{\sqrt{\sigma_{ii}\lambda_j}} = \frac{\sqrt{\lambda_j} a_{ji}}{\sqrt{\sigma_{ii}}}$$

만약 주성분이 상관계수 행렬에서 얻어졌다면 $\sigma_{ii} = 1$이므로

$$Corr(X_i, Z_j) = \sqrt{\lambda_j} a_{ji}$$

따라서 주성분을 만드는 경우 주성분의 표준 편차 $\sqrt{\lambda_j}$ 으로 나누어 주면 원래 자료와의 공분산이 주성분변환의 계수로 나오게 된다.

$$Cov\left(X_i, \frac{Z_j}{\sqrt{\lambda_j}}\right) = a_{ji}$$

6.3.3. 주성분 점수

자료의 변동을 설명하기 위하여 m개의 주성분이 필요하다면 각 표본의 개체에 대한 새로운 주성분의 값을 구할 수 있다. 이를 주성분 점수(Principal Component score)라고 한다.

p개의 변수를 가지는 확률벡터를 독립적으로 n개 표본추출했다고 하자. X_{ij}는 i번째 표본의 j번째 변수이라고 하면 i번째 개체의 관측벡터는 $\boldsymbol{X}_i$는 다음과 같다.

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})^t$$

j번째 변수의 표본 평균을 $\bar{X}_j = \sum_{i=1}^{n} X_{ij}/n$ 으로 정의하고 추출된 표본의 평균벡터를 $\bar{\boldsymbol{X}}$를 다음과 같다고 하자.

$$\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)^t$$

이제 각 관측값을 해당하는 변수의 평균으로 빼준 벡터 $\boldsymbol{Y}_i$ 를 고려하자.

$$\boldsymbol{Y}_i = \boldsymbol{X}_i - \bar{\boldsymbol{X}} = (X_{i1} - \bar{X}_1, X_{i2} - \bar{X}_2, \dots, X_{ip} - \bar{X}_p)^t$$

이제 i번째 개체의 j 번째 주성분 점수는 다음과 같이 정의된다.

$$Z_{ij} = \boldsymbol{a}_j^t \boldsymbol{Y}_i = \sum_{k=1}^{p} a_{jk} Y_{ik} = \sum_{k=1}^{p} a_{jk}(X_{ik} - \bar{X}_k) \tag{6.11}$$

주성분 점수는 각각 관측된 개체가 원래 p개의 변수를 가지고 있다면 주성분분석을 통해서 더 적은 수, $m < p$개의 새로운 변수로서 관측 변수의 개수를 축소했다고 생각하면 된다.

$$\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{im})^t$$

6.4. 예제: 올림픽 7종 경기 자료

올림픽 7종 경기 자료에 주성분분석을 적용해보자. 7종 경기의 종목은 100m 허들, 높이뛰기, 투포환, 200m 달리기, 멀리뛰기, 창던지기, 800m달리기이다. 순위는 혼성 경기 득점표에 따라 각 종목의 종합 득점으로 정한다.

올림픽 7종 경기 자료는 다음과 같이 불러올 수 있다. 25명의 선수들에 대한 7개의 종목의 기록과 총점(`score`)으로 구성된 자료이다. 마지막 변수 `score` 는 올림픽 위원회가 7개의 종목의 기록을 결합해서 만든 점수로서 메달의 순위를 결정한다.

```
library(HSAUR2)

df_1 <- heptathlon
dim(df_1)
```

```
[1] 25  8
```

```
colnames(df_1)
```

```
[1] "hurdles"  "highjump" "shot"     "run200m"  "longjump" "javelin"  "run800m"
[8] "score"
```

```
head(df_1)
```

```
                   hurdles highjump  shot run200m longjump javelin run800m
Joyner-Kersee (USA)   12.69     1.86 15.80   22.56     7.27   45.66  128.51
John (GDR)            12.85     1.80 16.23   23.65     6.71   42.56  126.12
Behmer (GDR)          13.20     1.83 14.20   23.10     6.68   44.54  124.20
Sablovskaite (URS)    13.61     1.80 15.23   23.92     6.25   42.78  132.24
Choubenkova (URS)     13.51     1.74 14.76   23.93     6.32   47.46  127.90
Schulz (GDR)          13.75     1.83 13.50   24.65     6.33   42.82  125.79
                   score
Joyner-Kersee (USA)  7291
John (GDR)           6897
Behmer (GDR)         6858
Sablovskaite (URS)   6540
Choubenkova (URS)    6540
Schulz (GDR)         6411
```

먼저 7개의 자료의 값이 크기가 크면 좋은 기록이 나타내도록 변환을 한다. 따라서 육상경기 변수는 각 관측값을 그 변수의 최대값에서 빼준다. 또한 파푸아뉴기니아 선수가 너무 실력이 떨어져서 이상치로 분류하여 분석에서 제외한다.

```
df_1$hurdles <- with(df_1, max(hurdles)-hurdles)
df_1$run200m <- with(df_1, max(run200m)-run200m)
df_1$run800m <- with(df_1, max(run800m)-run800m)

df_1 <- df_1 %>%
            filter(!str_detect(rownames(.), "PNG"))
dim(df_1)
```

```
[1] 24  8
```

```
head(df_1)
```

```
                   hurdles highjump  shot run200m longjump javelin run800m
Joyner-Kersee (USA)    3.73     1.86 15.80    4.05     7.27   45.66   34.92
John (GDR)             3.57     1.80 16.23    2.96     6.71   42.56   37.31
Behmer (GDR)           3.22     1.83 14.20    3.51     6.68   44.54   39.23
Sablovskaite (URS)     2.81     1.80 15.23    2.69     6.25   42.78   31.19
```

```
Choubenkova (URS)      2.91     1.74 14.76    2.68     6.32   47.46   35.53
Schulz (GDR)           2.67     1.83 13.50    1.96     6.33   42.82   37.64
                     score
Joyner-Kersee (USA)  7291
John (GDR)           6897
Behmer (GDR)         6858
Sablovskaite (URS)   6540
Choubenkova (URS)    6540
Schulz (GDR)         6411
```

먼저 7개 변수의 상관관계를 보자

```
df_1_t <- df_1 %>%
          select(-score)

cor(df_1_t)
```

```
           hurdles  highjump      shot   run200m  longjump   javelin   run800m
hurdles  1.0000000 0.5817409 0.7666860 0.8300371 0.8893472 0.3324779 0.5587794
highjump 0.5817409 1.0000000 0.4646854 0.3909024 0.6626910 0.3480793 0.1523350
shot     0.7666860 0.4646854 1.0000000 0.6694330 0.7840380 0.3430333 0.4082925
run200m  0.8300371 0.3909024 0.6694330 1.0000000 0.8106176 0.4707969 0.5731902
longjump 0.8893472 0.6626910 0.7840380 0.8106176 1.0000000 0.2870826 0.5233809
javelin  0.3324779 0.3480793 0.3430333 0.4707969 0.2870826 1.0000000 0.2559348
run800m  0.5587794 0.1523350 0.4082925 0.5731902 0.5233809 0.2559348 1.0000000
```

주성분 분석은 함수 `princomp` 를 사용한다. 맨 마지막 8번째 변수 `score` 를 제외한 7개의 변수로 주성분분석을 실시하자. 함수 `princomp` 에서 선택문 `cor = TRUE` 는 상관계수행렬로 주성분분석을 실행하라는 선택이며 함수 `summary` 에서 `loadings = TRUE` 는 주성분의 계수를 보여주라는 것이다.

```
pca1 <- princomp(df_1_t, cor = TRUE)

summary(pca1,loadings = TRUE)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.0793370 0.9481532 0.9109016 0.68319667 0.54618878
Proportion of Variance 0.6176632 0.1284278 0.1185345 0.06667967 0.04261745
Cumulative Proportion  0.6176632 0.7460909 0.8646255 0.93130515 0.97392260
                           Comp.6      Comp.7
Standard deviation     0.33745486 0.262042024
Proportion of Variance 0.01626797 0.009809432
```

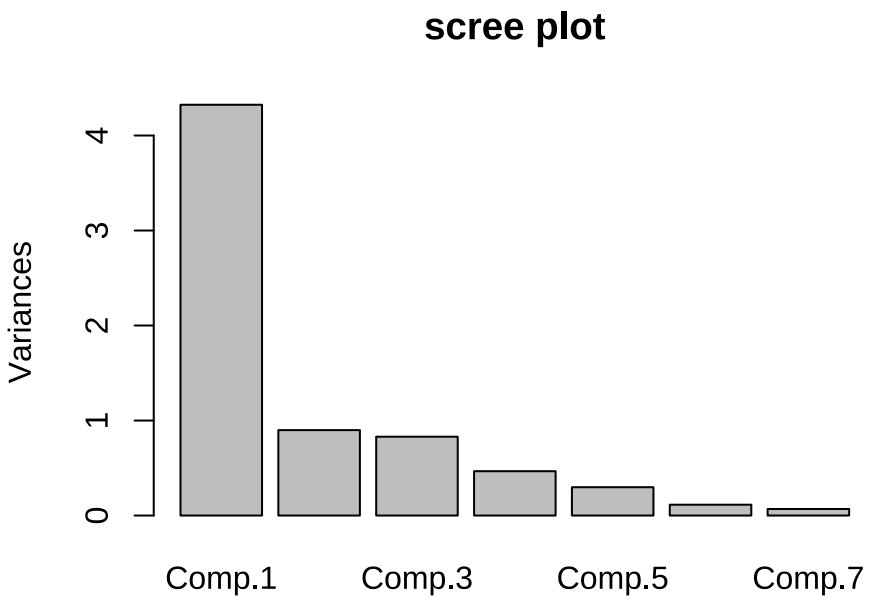
```
Cumulative Proportion  0.99019057 1.000000000

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
hurdles   0.450         0.174         0.199  0.847
highjump  0.315 -0.651  0.209 -0.557                -0.332
shot      0.402         0.153  0.548 -0.672         -0.229
run200m   0.427  0.185 -0.130  0.231  0.618 -0.333 -0.470
longjump  0.451         0.270         0.122 -0.383  0.749
javelin   0.242 -0.326 -0.881                       0.211
run800m   0.303  0.657 -0.193 -0.574 -0.319
```

주성분 분석의 결과를 다음과 같이 해석할 수 있다.

- 주성분 2개를 선택하면 총변동의 76 % 를 설명하고 3개를 선택하면 86 % 를 설명한다.
  - `Proportion of Variance` 은 식 6.9 에서 정의된 P_i 의 값을 나타낸다.
  - `Cumulative Proportion` 식 6.10 의 P_i 이 누적합을 나타낸다.
- 먼저, 계수의 값이 없는 이유는 상대적으로 작은 값은 제시하지 않아서 이다
- 첫번째 주성분(`Comp.1`)의 계수를 보면 각 종목 점수의 가중평균과 유사하다.
- 두 번째 주성분(`Comp.2`)은 800달리기와 높이뛰기의 차이로 해석된다(각각의 계수가 다른 변수에 비해 절상대적으로 크고 부호가 반대이다).
- 3번째 주성분(`Comp.3`)은 창던지기(javelin)의 능력을 나타내는것으로 해석된다.

주성분 개수의 선택에 이용되는 scree plot은 다음과 같다. 막대의 높이는 주성분의 분산, 즉 고유치와 비례한다

```
plot(pca1, main = "scree plot")
```



25명의 선수에 대한 식 6.11 에 의한 주성분 점수를 구할 수 있고 첫번째와 두번째 주성분 점수를 산점도로 그려보았다.

```
head(pca1$scores)
```

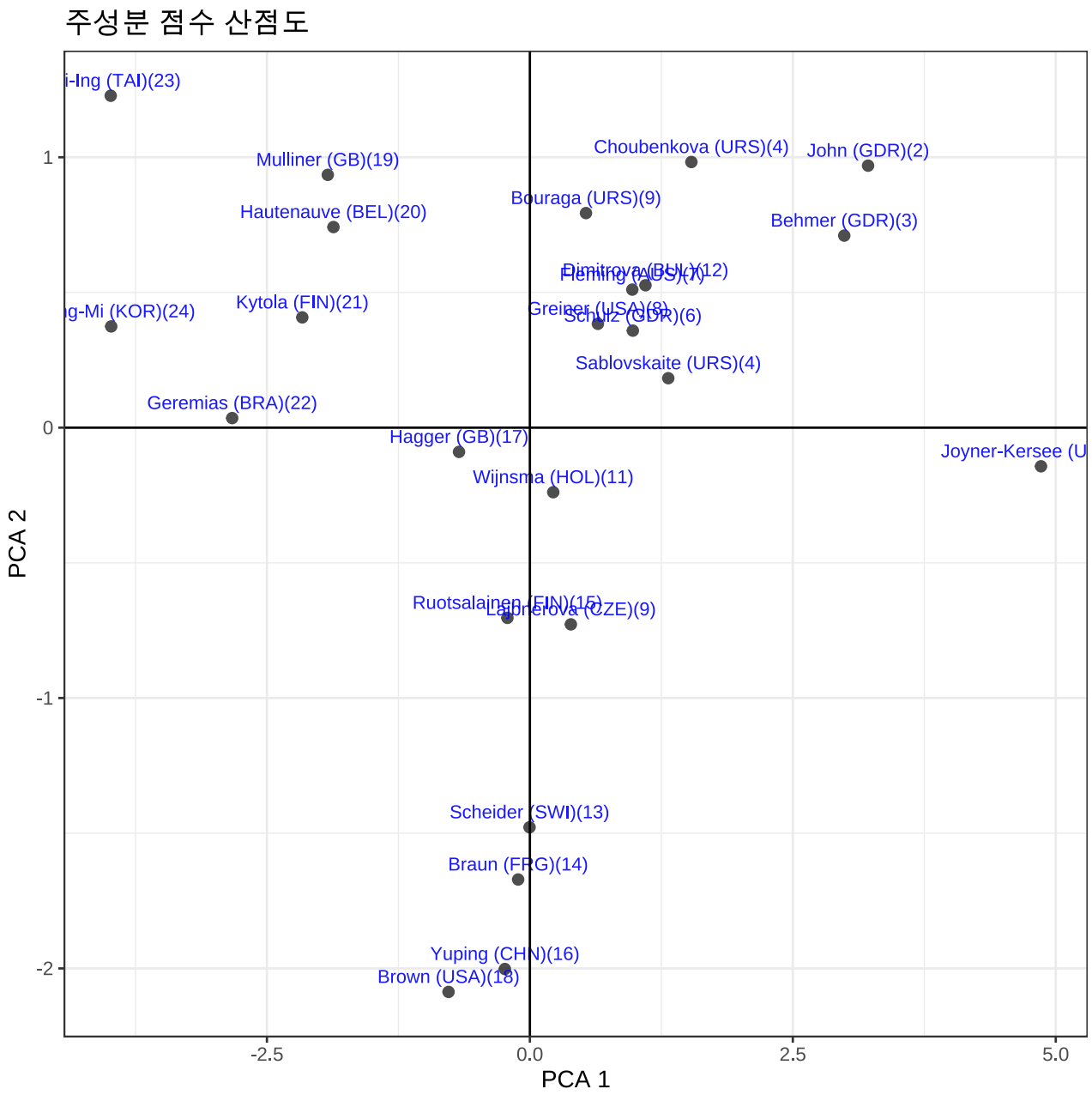
```
                      Comp.1     Comp.2       Comp.3     Comp.4     Comp.5
Joyner-Kersee (USA) 4.8598544 -0.1428696  0.006170444  0.2997271  0.3696153
John (GDR)          3.2156489  0.9689924  0.249166030  0.5609829 -0.7698539
Behmer (GDR)        2.9891207  0.7102977 -0.635677626 -0.5666763  0.1944444
Sablovskaite (URS)  1.3158405  0.1828572 -0.256022940  0.6508784 -0.6166041
Choubenkova (URS)   1.5357870  0.9824589 -1.818885157  0.8008982 -0.6023826
Schulz (GDR)        0.9790817  0.3587703 -0.421970960 -1.1374969 -0.7302135
                         Comp.6      Comp.7
Joyner-Kersee (USA) -0.27632076  0.48611032
John (GDR)           0.38582368  0.05283965
Behmer (GDR)        -0.26334765 -0.11292729
Sablovskaite (URS)  -0.22039763 -0.54216684
Choubenkova (URS)    0.08186703  0.30728837
Schulz (GDR)        -0.25984050 -0.03921361
```

```
df_pca <- as.data.frame(pca1$scores)

df_pca <- df_pca %>%
  rownames_to_column("label") %>%
```

```
  left_join(df_1 %>% rownames_to_column("label") %>% select(label, score), by="label") %>%
  mutate(rank = as.integer(rank(-score))) %>%
  mutate(label = paste0(label, "(", rank, ")"))


head(df_pca)
```

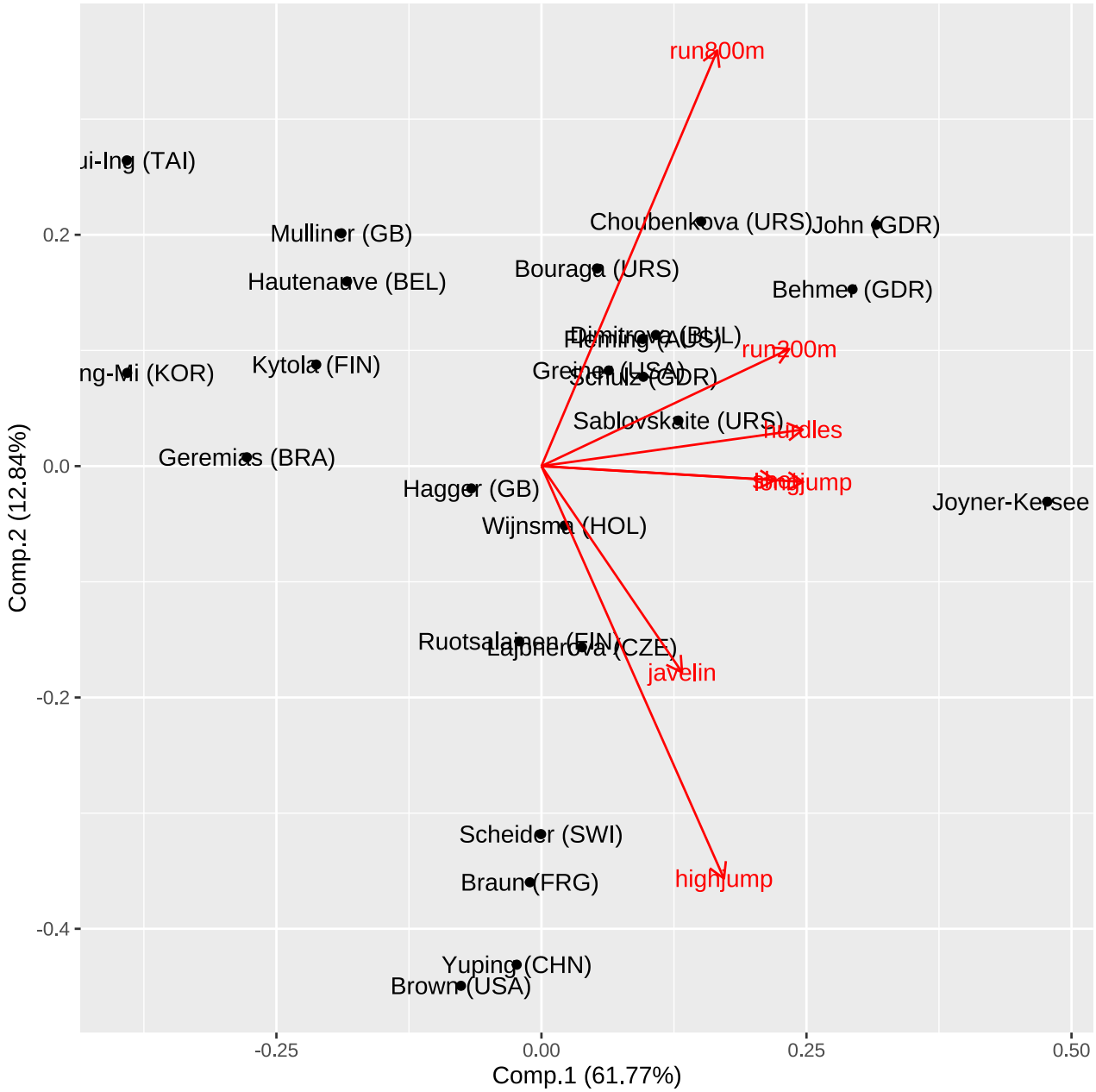
```
                  label    Comp.1     Comp.2       Comp.3     Comp.4
1 Joyner-Kersee (USA)(1) 4.8598544 -0.1428696  0.006170444  0.2997271
2          John (GDR)(2) 3.2156489  0.9689924  0.249166030  0.5609829
3        Behmer (GDR)(3) 2.9891207  0.7102977 -0.635677626 -0.5666763
4  Sablovskaite (URS)(4) 1.3158405  0.1828572 -0.256022940  0.6508784
5   Choubenkova (URS)(4) 1.5357870  0.9824589 -1.818885157  0.8008982
6        Schulz (GDR)(6) 0.9790817  0.3587703 -0.421970960 -1.1374969
      Comp.5      Comp.6      Comp.7 score rank
1  0.3696153 -0.27632076  0.48611032  7291    1
2 -0.7698539  0.38582368  0.05283965  6897    2
3  0.1944444 -0.26334765 -0.11292729  6858    3
4 -0.6166041 -0.22039763 -0.54216684  6540    4
5 -0.6023826  0.08186703  0.30728837  6540    4
6 -0.7302135 -0.25984050 -0.03921361  6411    6
```

```
df_pca%>%
  ggplot(aes(x=Comp.1, y=Comp.2 , label=label, size=score)) +
  geom_point(size= 2, alpha = 0.7) +
  geom_text(size= 3 ,vjust = -0.7, color = "blue", alpha=0.9) +
  geom_hline(yintercept=0) +
  geom_vline(xintercept=0) +
  theme_bw() +
  labs(x="PCA 1", y="PCA 2", title = "주성분 점수 산점도")
```



Biplot 은 25개의 개체에 대하여 2개의 주성분 점수와 그 방향을 이차원 공간에 표시한 그림으로 주성분의 구조를 이해하는데 큰 도움이 된다. Biplot 은 패키지 `ggfortify` 의 `autoplot` 함수를 이용하면 다음과 같이 그릴 수 있다.

```
autoplot(pca1, label = TRUE, loadings = TRUE, loadings.label = TRUE)
```



화살표로 표시된 각 변수에 대한 방향 벡터는 2개의 주성분 계수들 중 각 변수에 해당하는 두 계수(`Comp.1`, `Comp.2`) 벡터의 방향이다. 위의 Biplot에서 다음과 같은 사실을 알 수 있다.

- 각 변수에 대한 벡터의 방향은 두 주성분의 축에서 더 평행한 방향으로 주성분 점수에 기여도가 높다.
- 각 변수에 대한 벡터의 각도는 상관성을 나타내며 그 각이 작을수록 상관관계가 크다.
- 첫번째 주성분은 `hurdle`, `longjump`, `200m`가 변동을 주로 설명한다.
  - 선수 Joyner가 오른쪽 끝에 위치하는 것은 첫번째 주성분 점수가 가장 크고 `hurdle`, `longjump`, `200m`에서 좋은 점수를 낸것을 알려준다.
- 두번째 주성분은 `800m`와 `highjump`가 변동을 주로 설명하며 두 변수의 상관관계가 상대적으로 작으므로 방향이 반대인 것을 알 수 있다.

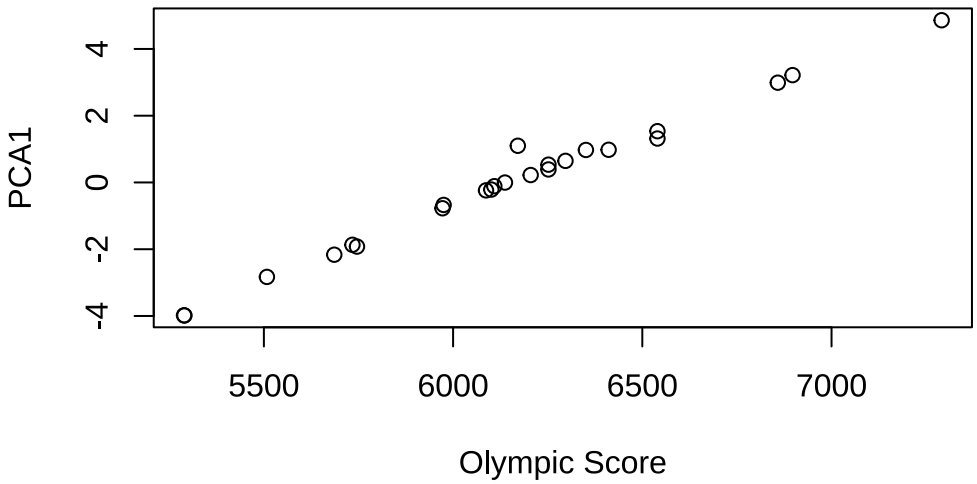
- 800m와 highjump 벡터 방향이 서로 반대이며 각 벡터 방향에 가까운 선수들이 해당 종목의 점수가 높음을 알 수 있다.
- 예를 들어 선수 Yuping 은 highjump 점수가 높으므로 highjump 방향인 아래쪽에 위치하고 있다.

마지막으로 첫번째 주성분 점수와 올림픽 위원회가 구한 총점과의 관계를 보자. 두 값이 매우 강한 상관관계를 나타낸다.

```
cor(df_1$score, df_pca$`Comp.1`)
```

```
[1] 0.9931168
```

```
plot(df_1$score, df_pca$`Comp.1`, xlab="Olympic Score", ylab="PCA1")
```



6.5. 예제: 포도주 자료

포도주 자료는 먼저 1978년에 이탈리아 피에몬테(Piedmont) 지역에서 재배된 세 가지 포도 품종(cultivar) 와인의 화학 성분을 측정한 자료이다. 전체 178 개의 와인이 3개의 그룹으로 나누어져 있다. 변수 Class 는 포도의 품종을 나타내는 범주형 변수이다.

Class	포도 품종(지역)
1	Barolo (바롤로)
2	Grignolino (그리뇰리노)
3	Barbera (바르베라)

포도주 자료는 사용와인의 화학 조성을 기반으로 주요 성분 간 상관관계를 분석할 수 있으며 포도 품종 간 차이 비교하는데 사용된다. 차원 축소(PCA) 및 군집 분석에 자주 사용되는 자료이다.

13개의 변수에 대한 설명은 다음과 같다.

변수명	설명
Class	포도 품종 (1, 2, 3 세 가지 품종 구분)
Alcohol	알코올 함량 (%)
Malic_Acid	사과산 함량 – 와인의 신맛을 결정하는 유기산
Ash	회분 – 무기질 함량 지표
Alcalinity_of_Ash	회분의 알칼리도 – 산-알칼리 균형 지표
Magnesium	마그네슘 함량 (mg/L)
Total_Phenols	총 페놀 성분 – 향, 맛, 항산화 효과에 기여
Flavanoids	플라보노이드 성분 – 떫은맛, 색, 향과 관련
Nonflavanoid_Phenols	비플라보노이드 페놀 성분
Proanthocyanins	프로안토시아닌 – 탄닌 전구체, 색과 떫은맛 영향
Color_Intensity	색 농도 – 와인의 색상이 얼마나 짙은지
Hue	색조 – 적색 대비 황색 비율
OD280_OD315	280nm/315nm 흡광도 비율 – 페놀 함량 및 품질 지표
Proline	프롤린 함량 – 아미노산, 와인 향과 숙성에 중요한 역할

6.5.1. 자료 불러오기

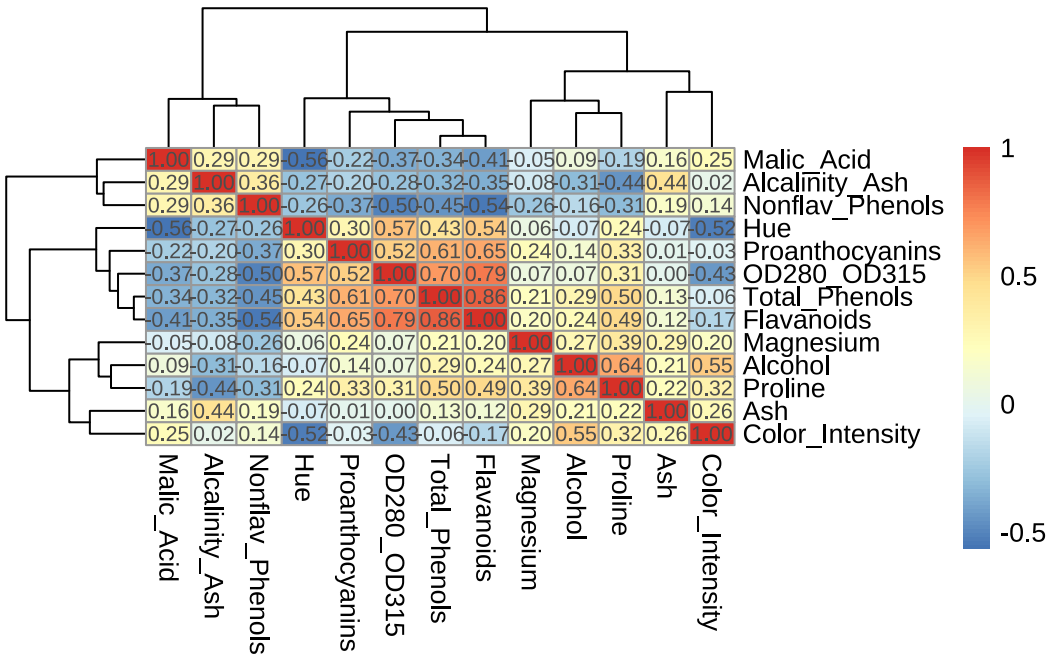
```
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"
wine <- read.csv(url, header = FALSE)

# Add column names from UCI description
colnames(wine) <- c("Class","Alcohol","Malic_Acid","Ash","Alcalinity_Ash","Magnesium",
```

```
                    "Total_Phenols","Flavanoids","Nonflav_Phenols","Proanthocyanins",
                    "Color_Intensity","Hue","OD280_OD315","Proline")
head(wine)
```

```
  Class Alcohol Malic_Acid  Ash Alcalinity_Ash Magnesium Total_Phenols
1     1   14.23       1.71 2.43           15.6       127          2.80
2     1   13.20       1.78 2.14           11.2       100          2.65
3     1   13.16       2.36 2.67           18.6       101          2.80
4     1   14.37       1.95 2.50           16.8       113          3.85
5     1   13.24       2.59 2.87           21.0       118          2.80
6     1   14.20       1.76 2.45           15.2       112          3.27
  Flavanoids Nonflav_Phenols Proanthocyanins Color_Intensity  Hue OD280_OD315
1       3.06            0.28            2.29            5.64 1.04        3.92
2       2.76            0.26            1.28            4.38 1.05        3.40
3       3.24            0.30            2.81            5.68 1.03        3.17
4       3.49            0.24            2.18            7.80 0.86        3.45
5       2.69            0.39            1.82            4.32 1.04        2.93
6       3.39            0.34            1.97            6.75 1.05        2.85
  Proline
1    1065
2    1050
3    1185
4    1480
5     735
6    1450
```

먼저 품종(Class)을 제외한 13개의 변수의 상관계수 행렬을 시각화 해보자.

```
## 품종 제외한 변수 선택
df_2 <- wine %>% select(-Class)

## 상관계수 행렬 계산
cor_mat <- cor(df_2, use="pairwise.complete.obs")

# 히트맵 그리기
pheatmap(cor_mat,  display_numbers = TRUE, number_format = "%.2f")
```



상관계수 행렬을 보면 변수들이 3개의 그룹으로 나누어져 있다것을 알 수 있다.

6.5.2. 주성분 분석

이제 `wine` 자료에 대하여 주성분 분석을 적용해 보자.

```
pca_wine <- princomp(df_2, cor = TRUE)
summary(pca_wine, loadings = TRUE)
```

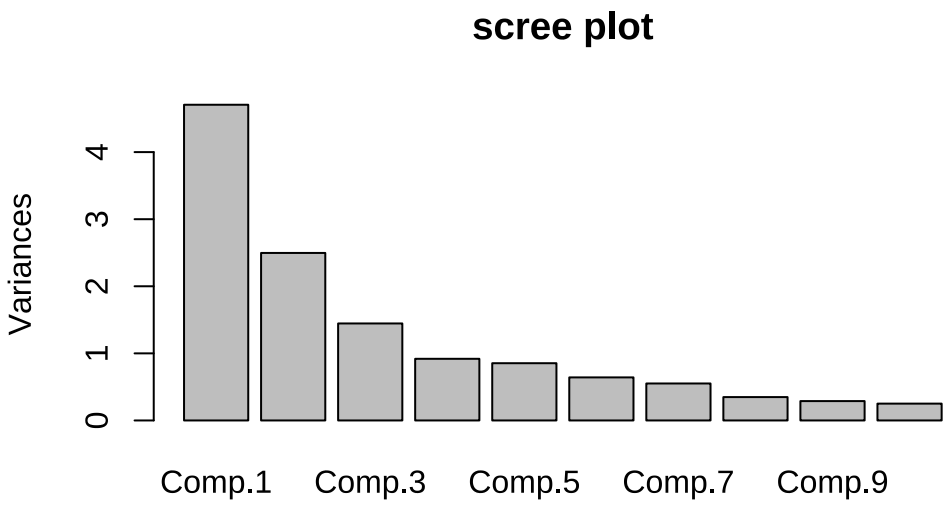
```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     2.1692972 1.5801816 1.2025273 0.9586313 0.92370351
Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294
Cumulative Proportion  0.3619885 0.5540634 0.6652997 0.7359900 0.80162293
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.80103498 0.74231281 0.59033665 0.53747553 0.50090167
Proportion of Variance 0.04935823 0.04238679 0.02680749 0.02222153 0.01930019
Cumulative Proportion  0.85098116 0.89336795 0.92017544 0.94239698 0.96169717
                          Comp.11    Comp.12     Comp.13
Standard deviation     0.47517222 0.41081655 0.321524394
Proportion of Variance 0.01736836 0.01298233 0.007952149
Cumulative Proportion  0.97906553 0.99204785 1.000000000

Loadings:
               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Alcohol         0.144  0.484  0.207         0.266  0.214         0.396  0.509
```

```
Malic_Acid      -0.245  0.225        -0.537         0.537 -0.421
Ash                     0.316 -0.626  0.214  0.143  0.154  0.149 -0.170 -0.308
Alcalinity_Ash  -0.239        -0.612                -0.101  0.287  0.428  0.200
Magnesium        0.142  0.300 -0.131  0.352 -0.727         -0.323 -0.156  0.271
Total_Phenols    0.395        -0.146 -0.198  0.149                -0.406  0.286
Flavanoids       0.423        -0.151 -0.152  0.109                -0.187
Nonflav_Phenols -0.299        -0.170  0.203  0.501 -0.259 -0.595 -0.233  0.196
Proanthocyanins  0.313        -0.149 -0.399 -0.137 -0.534 -0.372  0.368 -0.209
Color_Intensity         0.530  0.137                -0.419  0.228
Hue              0.297 -0.279         0.428  0.174  0.106 -0.232  0.437
OD280_OD315      0.376 -0.164 -0.166 -0.184  0.101  0.266                0.137
Proline          0.287  0.365  0.127  0.232  0.158  0.120         0.120 -0.576
                Comp.10 Comp.11 Comp.12 Comp.13
Alcohol          0.212   0.226   0.266
Malic_Acid      -0.309          -0.122
Ash                      0.499          -0.141
Alcalinity_Ash          -0.479
Magnesium
Total_Phenols   -0.320  -0.304   0.304  -0.464
Flavanoids      -0.163                  0.832
Nonflav_Phenols  0.216  -0.117          0.114
Proanthocyanins  0.134   0.237         -0.117
Color_Intensity -0.291          -0.604
Hue             -0.522          -0.259
OD280_OD315      0.524          -0.601  -0.157
Proline          0.162  -0.539
```

- 첫 번째 주성분의 계수를 보면 `Flavanoids` , `Total_Phenols`, `OD280_OD315`, `Proanthocyanins` 의 계수값이 양수로 다른 변수들에 비하여 상재적으로 크게 나타난 것을 알 수 있다. 또한 앞의 상관관계 분석에서도 4개 변수들의 상관관계가 높다는 것을 알 수 있다. 4개의 변수들은 맛의 풍부함 또는 맛의 강도와 관련된 변수들이다.
- 두번째 주성분의 계수를 보면 `Color_Intensity` , `Alcohol`, `Ash`, `Hue` 의 계수값이 양수로서 다른 변수들에 비하여 상대적으로 크게 나타난 것을 알 수 있다. 이러한 변수들는 서로 상관관계가 높다는 것을 알 수 있다. 이 변수들 중 `Color_Intensity` 와 `Hue` 는 색상에 관련된 특성을 가지고 있다.

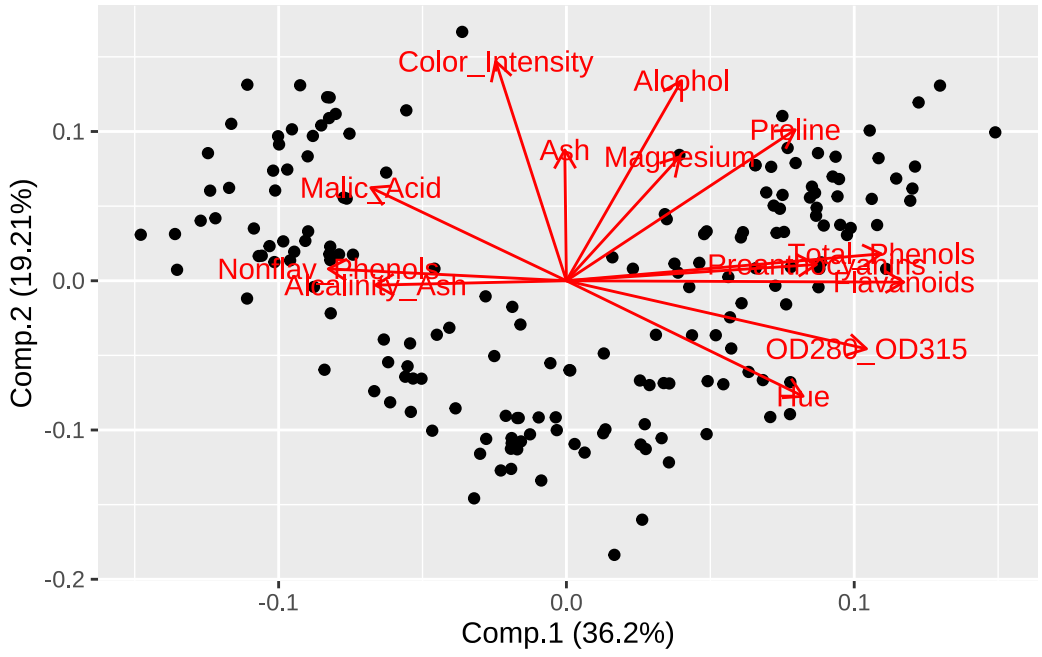
자료의 변동을 설명하는 비율을 보기 위해서 scree plot 을 만들어 보자. 앞의 결과에서 보았듯이 두 개의 주성분이 55%의 변동을 설명한다

```
plot(pca_wine, main = "scree plot")
```



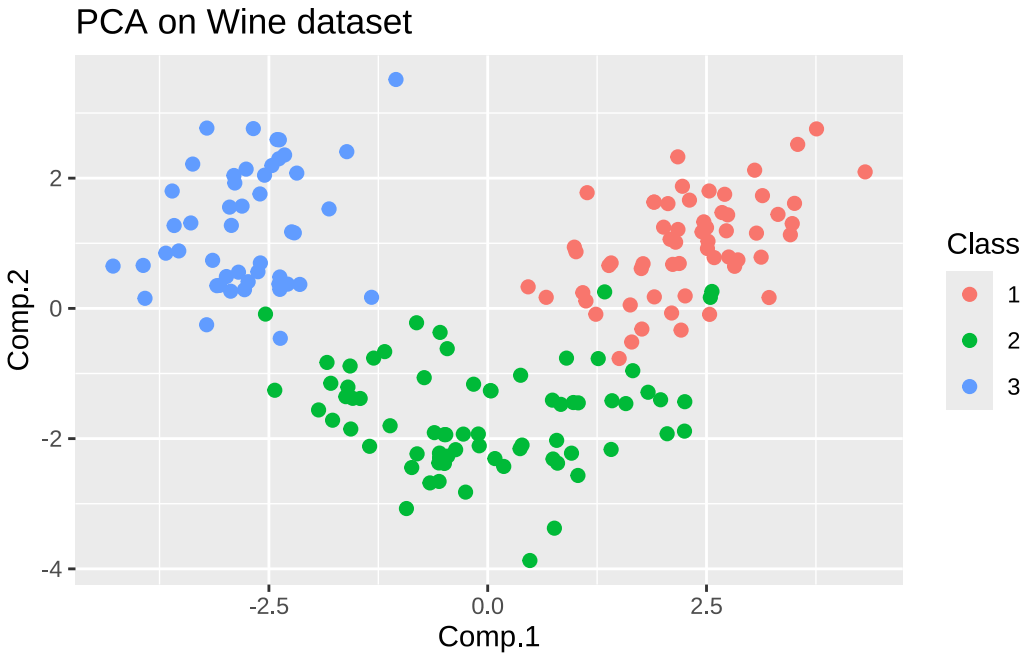
Biplot 은 178개의 포도주 샘플에 대하여 2개의 주성분 점수와 그 방향을 이차원 공간에 표시한 그림으로 주성분의 구조를 이해하는데 큰 도움이 된다. 아래 그림에서 보듯이 첫 번째 주성분과 두 번째 주성분을 구성하는 변수들의 역할을 시각화하여 볼 수 있다.

```
autoplot(pca_wine, label = FALSE, loadings = TRUE, loadings.label = TRUE)
```



이제 마지막으로 주성분 분석을 통해 얻은 2개의 주성분 점수를 이용하여 포도 품종 간의 차이를 비교해 보자. 14개의 변수가 아닌 2개의 주성분으로 차원을 축소했음에도 불구하고 3개의 품종이 잘 구분되는 것을 알 수 있다.

```r
#- 주성분 점수 데이터프레임 생성
df_pca <- as.data.frame(pca_wine$scores)

#- 품종 정보를 데이터프레임에 추가
df_pca$Class <- factor(wine$Class)

#- 주성분 점수 산점도 그리기
ggplot(df_pca, aes(Comp.1, Comp.2, color = Class)) +
  geom_point(size = 2) +
  labs(title = "PCA on Wine dataset")
```



7. 정준상관분석

```
library(tidyverse)
library(here)
library(knitr)
library(CCA)

#아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
#font_add_google("Nanum Pen Script", "gl")
font_add_google(name = "Noto Sans KR", family = "noto")
showtext_auto()
```

7.1. 상관계수

두 개의 확률변수의 상관계수(Correlation Cefficient)는 두 변수의 선형 관계의 정도를 나타내는 측도이다. 두 개의 확률변수 X_1과 X_2 의 상관계수 ρ 는 다음과 같이 정의된다.

$$\rho = \rho(X_1, X_2) = cor(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}} \tag{7.1}$$

상관계수 ρ는 -1과 1 사이의 값을 가지며 상관계수가 0 이면 두 확률변수의 선형관계는 존재하지 않는다. 상관계수가 1 에 가까울수록 두 변수는 양의 선형관계가 강해지며 반대로 -1 에 가까울수록 두 변수는 음의 선형관계가 강해진다. 두 변수의 상관계수가 0이라고 해서 관계(relation)가 없다고 단정할 수 없다. 왜냐하면 상관계수는 두 변수의 선형관계(linear relationship)만을 나타내는 측도이고 비선형 관계 등 다른 특별한 관계를 반영하지는 못한다.

두 개의 변수에 대한 n개의 독립표본 $(X_{11}, X_{12}), (X_{21}, X_{22}), \dots, (X_{n1}, X_{n2})$이 주어지면 표본으로 부터 모집단의 상관계수를 추정할 수 있는 표본 상관계수 $\hat{\rho}$ 을 다음과 같이 계산할 수 있다.

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2(X_{i2} - \bar{X}_2)^2}} \tag{7.2}$$

여기서 $\bar{X}_1$과 $\bar{X}_2$는 각각 첫 번째 변수의 표본 $X_{11}, X_{21}, \dots X_{n1}$과 두 번째 변수의 표본 $X_{12}, X_{22}, \dots X_{n2}$의 평균이다.

Example 7.1 (스위스의 47개 주 자료). R 패키지에 내장된 swiss 자료는 1988년 스위스의 47개 주에 대한 출산율과 사회경제변수를 모아놓은 자료이다($n = 47$). 6개의 변수에 대한 설명과 자료의 일부는 다음과 같다.

- Fertility: I_g, common standardized fertility measure
- Agriculture: % of males involved in agriculture as occupation
- Examination: % draftees receiving highest mark on army examination
- Education: % education beyond primary school for draftees.
- Catholic: % catholic (as opposed to protestant).
- Infant.Mortality: live births who live less than 1 year.

```
head(swiss)
```

```
             Fertility Agriculture Examination Education Catholic
Courtelary        80.2        17.0          15        12     9.96
Delemont          83.1        45.1           6         9    84.84
Franches-Mnt      92.5        39.7           5         5    93.40
Moutier           85.8        36.5          12         7    33.77
Neuveville        76.9        43.5          17        15     5.16
Porrentruy        76.1        35.3           9         7    90.57
             Infant.Mortality
Courtelary               22.2
Delemont                 22.2
Franches-Mnt             20.2
Moutier                  20.3
Neuveville               20.6
Porrentruy               26.6
```

자료에서 두개의 변수에 대한 상관계수는 cor 함수로 다음과 같이 계산할 수 있고 더 나아가 데이터프레임 swiss에 있는 모든 변수들에 대한 상관계수행렬도 동시에 계산할 수 있다.

```
cor(swiss)
```

```
                 Fertility Agriculture Examination   Education   Catholic
Fertility        1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
Agriculture      0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
Examination     -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
Education       -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
Catholic         0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
Infant.Mortality 0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
                 Infant.Mortality
Fertility              0.41655603
Agriculture           -0.06085861
```

```
Examination          -0.11402160
Education            -0.09932185
Catholic              0.17549591
Infant.Mortality      1.00000000
```

7.2. 다중상관계수

상관계수는 두 개의 확률변수에 대한 선형관계를 나타내는 측도이다. 두 개의 변수에 대한 측도인 상관계수를 를 두 개의 확률벡터에 대한 관계를 나타내는 측도로 확장할 수 있다.

이렇게 두 개의 확률벡터에 대한 상관관계를 나타내는 측도를 정준상관계수(Canonical Correlation Coeffcient)라고 한다. 이 절에서는 두 개의 확률벡터에 대한 관계를 측정하는 정준상관계수를 정의하기 전에 하나의 확률변수와 여러 개의 변수를 포함하는 확률벡터의 관계를 나타내는 다중상관계수(Multiple Correlation Coefficient)을 먼저 정의하고자 한다.

확률벡터 $\boldsymbol{X}$를 p개의 확률변수로 이루어졌다고 하고 하나의 확률변수 X_1 (편의상 첫 번째 확률변수를 선택하였다)와 나머지 $p-1$개의 변수로 구성된 확률벡터를 $\boldsymbol{X}_*$라고 하자.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} X_1 \\ \boldsymbol{X}_* \end{bmatrix}, \quad Cov(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12}^t \\ \boldsymbol{\sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

위의 식에서 σ_{11}은 X_1의 분산, $\boldsymbol{\sigma}_{12}$은 X_1과 $\boldsymbol{X}_*$의 $(p-1)$-차원 공분산 벡터, $\boldsymbol{\Sigma}_{22}$는 $\boldsymbol{X}_*$의 $(p-1)\times(p-1)$-차원 공분산행렬이다.

하나의 확률변수와 여러 개의 확률변수로 구성된 확률벡터간의 선형관계는 X_1과 $\boldsymbol{X}_*$에 포함된 각각의 확률변수에 대하여 각각 $p-1$개의 상관계수를 구하여 따로 따로 파악할 수 있다.

$$\rho_{12} = cor(X_1, X_2),\ \rho_{13} = cor(X_1, X_3), \dots,\ \rho_{1p} = cor(X_1, X_p)$$

하지만 이러한 관계는 각각 두 개의 변수들에 대한 관계로서 하나의 확률변수와 확률벡터의 관계를 하나의 측도로서 종합적으로 반영하는 것은 아니다. 이렇게 확률변수와 확률벡터의 관계를 하나의 측도로 나타내기 위하여 주성분분석이나 회귀분석과 유사하게 확률벡터 $\boldsymbol{X}_*$에 포함된 확률변수들의 선형조합을 생각하고 확률변수 X_1과 선형조합으로 만들어진 새로운 확률변수 $\boldsymbol{a}^t\boldsymbol{X}_*$의 상관관계를 생각해 보자.

$$cor(X_1, a_2X_2 + a_3X_3 + \cdots + a_pX_p) = cor(X_1, \boldsymbol{a}^t\boldsymbol{X}_*)$$

위의 식에서 선형조합으로 만들어진 새로운 확률변수와의 상관계수는 계수벡터 $\boldsymbol{a}$의 계수값에 따라 달라진다. 이렇게 무수히 많은 값이 가능한 경우 최대의 상관계수를 가지는 계수를 고려하는 것이 다중상관계수의 정의이다.

$$\rho(X_1, \boldsymbol{X}_*) = \max_{\boldsymbol{a}} cor(X_1, \boldsymbol{a}^t \boldsymbol{X}_*) \tag{7.3}$$

최대의 상관계수를 가지는 계수를 구하기 위하여 먼저 임의의 벡터 $\boldsymbol{a}$에 대하여 X_1과 $\boldsymbol{a}^t\boldsymbol{X}_*$의 상관계수를 유도해 보자. 먼저

$$\begin{aligned} Cov(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) &= E[(X_1 - E(X_1))[\boldsymbol{a}^t\boldsymbol{X}_* - E(\boldsymbol{a}^t\boldsymbol{X}_*)] \\ &= \boldsymbol{a}^t E[(X_1 - E(X_1))[\boldsymbol{X}_* - E(\boldsymbol{X}_*)] \\ &= \boldsymbol{a}^t\boldsymbol{\sigma}_{12} \\ Var(\boldsymbol{a}^t\boldsymbol{X}_*) &= \boldsymbol{a}^t Var(\boldsymbol{X}_*)\boldsymbol{a} \\ &= \boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a} \end{aligned}$$

따라서

$$cor(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) = \frac{Cov(X_1, \boldsymbol{a}^t\boldsymbol{X}_*)}{\sqrt{Var(X_1)Var(\boldsymbol{a}^t\boldsymbol{X}_*)}} = \frac{\boldsymbol{a}^t\boldsymbol{\sigma}_{12}}{\sqrt{\sigma_{11}(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a})}}$$

여기서 한 가지 유의해야할 점은 계수벡터 $\boldsymbol{a}$의 계수들의 부호만을 바꾸면 다중상관계수의 부호가 바뀐다는 점이다.

$$cor(X_1, -\boldsymbol{a}^t\boldsymbol{X}_*) = \frac{-\boldsymbol{a}^t\boldsymbol{\sigma}_{12}}{\sqrt{\sigma_{11}[(-\boldsymbol{a}^t)\boldsymbol{\Sigma}_{22}(-\boldsymbol{a})]}} = -cor(X_1, \boldsymbol{a}^t\boldsymbol{X}_*)$$

따라서 다중상관계수를 구하는 경우에는 그 부호에 관계없이 절대값이 가장 큰 경우를 고려해야 한다. 이러한 점을 고려하여 다중상관계수를 다음과 같이 자신의 제곱값을 통하여 유도할 수 있다. 이제 다중상관계수의 계수벡터를 구하기 위한 다중상관계수의 제곱에 대한 최대값을 구해보자.

$$\begin{aligned} cor^2(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) &= \frac{(\boldsymbol{a}^t\boldsymbol{\sigma}_{12})^2}{\sigma_{11}(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a})} \\ &= \frac{(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\sigma}_{12})^2}{\sigma_{11}(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a})} \\ &\le \frac{(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{a})(\boldsymbol{\sigma}_{12}^t\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\sigma}_{12})}{\sigma_{11}(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a})} \\ &= \frac{(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a})(\boldsymbol{\sigma}_{12}^t\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12})}{\sigma_{11}(\boldsymbol{a}^t\boldsymbol{\Sigma}_{22}\boldsymbol{a})} \\ &= \frac{\boldsymbol{\sigma}_{12}^t\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12}}{\sigma_{11}} \end{aligned}$$

위의 유도식에서 공분산 행렬 $\boldsymbol{\Sigma}$의 제곱근 행렬 $\boldsymbol{\Sigma}^{1/2}$는 양정치 행렬의 성질 식 C.15 을 이용하였다.

또한 위의 유도식에서 사용한 부등식은 코쉬-쉬바르쯔(Cauchy-Schwarz) 부등식을 적용한 결과이다 (식 5.18 참조)

$$(\boldsymbol{\alpha}^t\boldsymbol{\beta})^2 \le (\boldsymbol{\alpha}^t\boldsymbol{\alpha})(\boldsymbol{\beta}^t\boldsymbol{\beta})$$

또한 위의 유도식에서 부등식의 등식이 성립하는 경우는 임의의 상수 λ에 대하여 다음과 같은 관계가 성립하는 경우이다.

$$\Sigma_{22}^{1/2}\boldsymbol{a} = \lambda\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\sigma}_{12} \quad \rightarrow \quad \boldsymbol{a} = \lambda\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12} \tag{7.4}$$

이제 X_1과 $\boldsymbol{X}_*$의 다중상관계수는 다음과 같이 구해진다.

$$\rho(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) = \frac{(\boldsymbol{\sigma}_{12}^t\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12})^{1/2}}{(\sigma_{11})^{1/2}} \tag{7.5}$$

여기서 $\boldsymbol{a} = \lambda\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{12}$, $0 \le \rho(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) \le 1$

다중상관계수의 계산은 다음과 같이 상관계수들로 구할 수 있다.

$$\rho(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) = (\boldsymbol{\rho}_{12}^t\boldsymbol{R}_{22}^{-1}\boldsymbol{\rho}_{12})^{1/2}$$

여기서 $\boldsymbol{\rho}_{12}$는 X_1과 $\boldsymbol{X}_*$의 상관계수 벡터이며 $\boldsymbol{R}_{22}$는 $\boldsymbol{X}_*$의 상관계수행렬이다. 이러한 계산식의 유도는 아래와 같은 공분산 행렬과 상관계수 행렬의 관계(2차원 확률벡터의 예)로부터 유도할 수 있다.

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1/\sqrt{\sigma_{22}} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1/\sqrt{\sigma_{11}} & 0 \\ 0 & 1/\sqrt{\sigma_{22}} \end{bmatrix}$$

표본으로서 자료가 주어진 경우에는 식 7.5 에서 각각의 모수에 대하여 그 추정량을 구하면 표본 다중상관계수를 구할 수 있다.

$$\hat{\rho}(X_1, \boldsymbol{a}^t\boldsymbol{X}_*) = \frac{(\hat{\boldsymbol{\sigma}}_{12}^t\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\sigma}}_{12})^{1/2}}{(\hat{\sigma}_{11})^{1/2}} = (\hat{\boldsymbol{\rho}}_{12}^t\hat{\boldsymbol{R}}_{22}^{-1}\hat{\boldsymbol{\rho}}_{12})^{1/2}$$

`swiss` 자료에서 출산율(X_1)과 나머지 5개의 사회경제변수($\boldsymbol{X}_*$)의 표본 다중상관계수는 다음과 같이 구할 수 있다.

```
n<-dim(swiss)[1] ; p <- dim(swiss)[2]
dim(swiss)
```

```
[1] 47  6
```

```
R <- cor(swiss)
R12 <- matrix(R[2:p,1],p-1,1)
R12
```

```
           [,1]
[1,]  0.3530792
[2,] -0.6458827
[3,] -0.6637889
[4,]  0.4636847
[5,]  0.4165560
```

```
R22 <- matrix(R[2:p,2:p],p-1,p-1)
R22
```

```
            [,1]       [,2]        [,3]       [,4]        [,5]
[1,]  1.00000000 -0.6865422 -0.63952252  0.4010951 -0.06085861
[2,] -0.68654221  1.0000000  0.69841530 -0.5727418 -0.11402160
[3,] -0.63952252  0.6984153  1.00000000 -0.1538589 -0.09932185
[4,]  0.40109505 -0.5727418 -0.15385892  1.0000000  0.17549591
[5,] -0.06085861 -0.1140216 -0.09932185  0.1754959  1.00000000
```

```
mulcor <- sqrt(t(R12) %*% solve(R22) %*% R12)
mulcor
```

```
          [,1]
[1,] 0.8406753
```

참고로 표본 다중상관계수의 제곱값은 는 X_1을 종속변수, $\boldsymbol{X}_*$을 독립변수로 선형회귀직선(linear regression)을 적합하였을 경우 결정계수(coefficient of determination) R^2 와 같다.

$$\hat{\rho}^2(X_1, \boldsymbol{a}^t \boldsymbol{X}_*) = R^2$$

아래 R 프로그램의 결과에서 확인할 수 있다.

```
res <- lm(Fertility~ Agriculture + Examination + Education + Catholic + Infant.Mortality,data=
summary(res)
```

```
Call:
lm(formula = Fertility ~ Agriculture + Examination + Education +
    Catholic + Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

```
mulcor^2
```

```
         [,1]
[1,] 0.706735
```

7.3. 정준상관계수

7.3.1. 정준상관계수의 정의

이제 두 개의 확률벡터에 대한 상관관계를 나타내는 측도인 정준상관계수(Canonical Correlation Coeffcient)에 대하여 알아보자. 확률벡터 $\boldsymbol{X}$가 두 개의 확률벡터 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$로 나누어져 있다고 가정하자. 두 확률 벡터의 차원은 각각 p와 q 라고 하자 (편의상 $p \leq q$라고 가정한다.)

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \quad Cov(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^t & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

앞 절에서와 유사한 방법으로 각 확률벡터의 선형조합으로 만들어진 두 개의 새로운 확률변수를 이용하여 정준상관계수를 다음과 같이 정의한다.

$$\rho(\boldsymbol{X}_1, \boldsymbol{X}_2) = \max_{\boldsymbol{a},\boldsymbol{b}} \; cor(\boldsymbol{a}^t\boldsymbol{X}_1, \boldsymbol{b}^t\boldsymbol{X}_2) \tag{7.6}$$

앞절에서 공분산행렬을 구할 때 사용한 같은 방법을 적용하면

$$\begin{aligned} Cov(\boldsymbol{a}^t\boldsymbol{X}_1, \boldsymbol{b}^t\boldsymbol{X}_2) &= E[(\boldsymbol{a}^t\boldsymbol{X}_1 - E(\boldsymbol{a}^t\boldsymbol{X}_1)][\boldsymbol{b}^t\boldsymbol{X}_2 - E(\boldsymbol{b}^t\boldsymbol{X}_2)] \\ &= \boldsymbol{a}^t E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1)][\boldsymbol{X}_2 - E(\boldsymbol{X}_2)]^t\boldsymbol{b} \\ &= \boldsymbol{a}^t\boldsymbol{\Sigma}_{12}\boldsymbol{b} \end{aligned}$$

따라서

$$cor(\boldsymbol{a}^t\boldsymbol{X}_1, \boldsymbol{b}^t\boldsymbol{X}_2) = \frac{\boldsymbol{a}^t\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{(\boldsymbol{a}^t\boldsymbol{\Sigma}_{11}\boldsymbol{a})(\boldsymbol{b}^t\boldsymbol{\Sigma}_{22}\boldsymbol{b})}} = (\boldsymbol{a}^t\boldsymbol{\Sigma}_{11}\boldsymbol{a})^{-1/2}(\boldsymbol{a}^t\boldsymbol{\Sigma}_{12}\boldsymbol{b})(\boldsymbol{b}^t\boldsymbol{\Sigma}_{22}\boldsymbol{b})^{-1/2}$$

이제 새로운 두 변수의 상관계수를 최대로 하는 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$를 찾으면 정준상관계수가 구해진다. 여기서 정준상관계수행렬(canonical correlation matrix) $\boldsymbol{C}$을 다음과 같이 정의하자.

$$\boldsymbol{C} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} \tag{7.7}$$

또한 정준상관계수행렬 $\boldsymbol{C}$ 의 Sigular Value Decomposition(SVD)를 다음과 같이 고려하자 (SVD 에 대한 자세한 내용은 Section C.4 를 참조)

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^t \tag{7.8}$$

여기서 $\boldsymbol{U}$와 $\boldsymbol{V}$ 는 각각 차원이 $p \times p$, $q \times q$인 정규직교행렬(orthonormal matrix)이고 $\boldsymbol{S}$는 $p \times q$ 행렬로 대각원소는 $\boldsymbol{C}\boldsymbol{C}^t$ 의 고유값 λ_i의 제곱근을 대각원소로 하고 비대각원소는 0이다.

$$\boldsymbol{U}^t\boldsymbol{U} = \boldsymbol{I}_p, \quad \boldsymbol{V}^t\boldsymbol{V} = \boldsymbol{I}_q$$

만약 $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p$와 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_q$를 각각 $\boldsymbol{U}$와 $\boldsymbol{V}$의 정규직교벡터라고 놓으면 다음과 같이 표시할 수 있다.

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^t = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \ldots\ \boldsymbol{u}_p] \begin{bmatrix} \sqrt{\lambda_1} & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & \sqrt{\lambda_p} & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^t \\ \boldsymbol{v}_2^t \\ \vdots \\ \boldsymbol{v}_q^t \end{bmatrix}$$

여기서 참고로 $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p$는 $\boldsymbol{C}\boldsymbol{C}^t$ 의 고유벡터이고 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_q$는 $\boldsymbol{C}^t\boldsymbol{C}$ 의 고유벡터이다. 또한 두 행렬 $\boldsymbol{C}\boldsymbol{C}^t$와 $\boldsymbol{C}^t\boldsymbol{C}$는 0이 아닌 고유값이 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$로 같다.

$$\boldsymbol{C}\boldsymbol{C}^t = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t\boldsymbol{\Sigma}_{11}^{-1/2} \quad \boldsymbol{C}^t\boldsymbol{C} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{12}^t\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$$

SVD 분해에 대한 자세한 내용은 Section C.4 를 참조하자.

7.3.2. 정준상관계수의 유도

다시 정준상관계수의 정의에 따른 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$를 찾는 문제로 돌아가서 두 벡터 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$을 표준화한 후에 선형조합을 고려한다.

$$Z = \boldsymbol{a}^t \boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1), \quad W = \boldsymbol{b}^t \boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)$$

여기서 상관계수는 상수를 더하거나 빼도 변함이 없으므로 평균벡터를 빼는것은 영향이 없다. 또한 두 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$ 대신에 $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{a}$와 $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{b}$를 고려해도 어차피 두 개의 임의의 벡터이므로 정준상관계수의 정의에는 영향을 미치지 않는다.

$$\begin{aligned}
\rho(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \max_{\boldsymbol{a},\boldsymbol{b}} \; cor[\boldsymbol{a}^t \boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1), \boldsymbol{b}^t \boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)] \\
&= \max_{\boldsymbol{a}_*,\boldsymbol{b}_*} \; cor(\boldsymbol{a}_*^t \boldsymbol{X}_1, \boldsymbol{b}_*^t \boldsymbol{X}_2)
\end{aligned}$$

두 확률변수 Z와 W의 공분산과 분산은 다음과 같이 주어지고

$$\begin{aligned}
Cov(Z, W) &= Cov(\boldsymbol{a}^t \boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1), \boldsymbol{b}^t \boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)) \\
&= \boldsymbol{a}^t \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{b} \\
&= \boldsymbol{a}^t \boldsymbol{C} \boldsymbol{b} \\
Var(Z) &= \boldsymbol{a}^t \boldsymbol{a} \\
Var(W) &= \boldsymbol{b}^t \boldsymbol{b}
\end{aligned}$$

따라서

$$cor(Z, W) = \frac{\boldsymbol{a}^t \boldsymbol{C} \boldsymbol{b}}{\sqrt{(\boldsymbol{a}^t \boldsymbol{a})(\boldsymbol{b}^t \boldsymbol{b})}} = \frac{\boldsymbol{a}^t}{\sqrt{(\boldsymbol{a}^t \boldsymbol{a})}} \boldsymbol{C} \frac{\boldsymbol{b}}{\sqrt{(\boldsymbol{b}^t \boldsymbol{b})}}$$

위의 식에서 확률변수 Z와 W의 상관계수는 두 벡터 $\boldsymbol{a}/\sqrt{\boldsymbol{a}^t \boldsymbol{a}}$와 $\boldsymbol{b}/\sqrt{\boldsymbol{b}^t \boldsymbol{b}}$의 길이가 1이므로 두 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$의 길이를 1로 가정해도 무방하다 ($\boldsymbol{a}^t \boldsymbol{a} = \boldsymbol{b}^t \boldsymbol{b} = 1$). 따라서 다음과 같이 표시할 수 있다.

$$cor(Z, W) = \boldsymbol{a}^t \boldsymbol{C} \boldsymbol{b}, \quad \boldsymbol{a}^t \boldsymbol{a} = \boldsymbol{b}^t \boldsymbol{b} = 1 \tag{7.9}$$

여기서 벡터공간의 알려진 사실을 이용한다. 두 벡터 $\boldsymbol{a}$와 $\boldsymbol{b}$는 각각 p와 q-차원 벡터이므로 다음과 같은 두 벡터 $ $와 $\boldsymbol{\beta}$가 존재하고 다음과 같이 표시할 수 있다.

$$\begin{aligned}
\boldsymbol{a} &= \boldsymbol{U}\boldsymbol{\alpha} = \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2 + \cdots + \alpha_p \boldsymbol{u}_p \\
\boldsymbol{b} &= \boldsymbol{V}\boldsymbol{\beta} = \beta_1 \boldsymbol{v}_1 + \beta_2 \boldsymbol{v}_2 + \cdots + \beta_q \boldsymbol{v}_q
\end{aligned} \tag{7.10}$$

여기서 $\boldsymbol{\alpha}^t \boldsymbol{\alpha} = \boldsymbol{\beta}^t \boldsymbol{\beta} = 1$.

이제 식 7.9 에 식 7.10 를 대입하고 식 식 7.8 의 SVD를 이용하면

$$\begin{aligned} cor(Z,W) &= \boldsymbol{a}^t\boldsymbol{C}\boldsymbol{b} \\ &= \boldsymbol{\alpha}^t\boldsymbol{U}^t\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^t\boldsymbol{V}\boldsymbol{\beta} \\ &= \boldsymbol{\alpha}^t\boldsymbol{S}\boldsymbol{\beta} \\ &= \sum_{i=1}^{p}\alpha_i\beta_i\sqrt{\lambda_i} \\ &\le (\max_i \sqrt{\lambda_i})\sum_{i=1}^{p}\alpha_i\beta_i \\ &= \sqrt{\lambda_1}\sum_{i=1}^{p}\alpha_i\beta_i \end{aligned}$$

위의 식에서 $cor(Z,W)$가 상한값(upper bound)와 같아지려면 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$의 계수는 다음과 같은 조건일 때이다.

$$\alpha_1 = 1,\ \alpha_2 = \cdots = \alpha_p = 0, \quad \beta_1 = 1,\ \beta_2 = \cdots = \beta_p = 0$$

따라서 표준화된 두 벡터의 정준상관계수는 $\boldsymbol{C}\boldsymbol{C}^t$ 의 최대고유값의 제곱근이 되고 선형조합을 만드는 두 벡터 $\boldsymbol{a}$ 와 $\boldsymbol{b}$는 각각 $\boldsymbol{C}\boldsymbol{C}^t$ 와 $\boldsymbol{C}^t\boldsymbol{C}$ 의 첫번째 고유벡터가 된다.

$$\begin{aligned} \rho(\boldsymbol{X}_1,\boldsymbol{X}_2) &= Cov(Z,W) \\ &= \max_{\boldsymbol{a},\boldsymbol{b}}\ \ cor(\boldsymbol{a}^t\boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1-\boldsymbol{\mu}_1), \boldsymbol{b}^t\boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)) \\ &= cor(\boldsymbol{u}_1^t\boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1-\boldsymbol{\mu}_1), \boldsymbol{v}_1^t\boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)) \\ &= \sqrt{\lambda_1} \end{aligned}$$

여기서 $\boldsymbol{a} = \boldsymbol{u}_1,\ \boldsymbol{b} = \boldsymbol{v}_1$ 이다.

정준상관계수를 정의하고 유도하는 과정을 보면 주성분분석과 유사하게 서로 공분산이 0인 확률변수들를 만들고 대응하는 두 변수들간의 상관계수를 큰 순서대로 만들 수 있다. 예를 들어 $Z_2 = \boldsymbol{u}_2^t\boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1-\boldsymbol{\mu}_1)$는 $Z_1 = \boldsymbol{u}_1^t\boldsymbol{\Sigma}_{11}^{-1/2}(\boldsymbol{X}_1-\boldsymbol{\mu}_1)$과 공분산이 0이고 또한 $W_2 = \boldsymbol{v}_2^t\boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)$는 $W_1 = \boldsymbol{v}_1^t\boldsymbol{\Sigma}_{22}^{-1/2}(\boldsymbol{X}_2-\boldsymbol{\mu}_2)$과 공분산이 0 이다 (정규분포인경우 독립이다). 더 나아가 Z_2와 W_2의 상관계수는 각각 Z_1과 W_1의 공분산이 0인 변수들중에서 최대의 상관계수 $\sqrt{\lambda_2}$ 를 가지게 된다. 유사한 방법으로 Z_i와 W_i들을 정의할 수 있다.

7.3.3. 표본 정준상관계수

이제 모집단이 아닌 표본이 주어졌을 경우 표본 정준상관계수를 구하는 방법은 다음과 같다. 이번에도 $p \le q$를 가정하자.

확률벡터 $\boldsymbol{X}$의 표본 상관계수행렬을 두 벡터 $\boldsymbol{X}_1, \boldsymbol{X}_2$의 차원에 맞게 다음과 같이 표시하고 구한다.

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{12}^t & \boldsymbol{R}_{22} \end{bmatrix}$$

다음과 같은 두 행렬을 구한다.

$$\boldsymbol{E}_1 = \boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{12}^t$$

$$\boldsymbol{E}_2 = \boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{12}^t\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}$$

두 벡터 $\boldsymbol{X}_1, \boldsymbol{X}_2$의 표본 정준상관계수는 $\boldsymbol{E}_1$ 또는 $\boldsymbol{E}_2$의 최대 고유치의 제곱근이다. 참고로 $\boldsymbol{E}_1$ 와 $\boldsymbol{E}_2$은 이 아닌 고유값이 같다.

`swiss` 자료에서 출산율(Fertility)과 영아사망율(Infant.Mortality)을 하나의 벡터로 나머지 4개의 사회경제변수를 다른 벡터로 하여 표본 정준상관계수는 다음과 같이 구할 수 있다.

```
n <-dim(swiss)[1]
p <- 2
q <- 4
swiss0 <- swiss[,c(1,6,2,3,4,5)] # 순서를 바꾸는 작업
R <- cor(swiss0)
R11 <- matrix(R[1:p,1:p],p,p)
R22 <- matrix(R[(p+1):(p+q),(p+1):(p+q)],q,q)
R12 <- matrix(R[1:p,(p+1):(p+q)],p,q)
R11
```

```
          [,1]     [,2]
[1,] 1.000000 0.416556
[2,] 0.416556 1.000000
```

```
R22
```

```
           [,1]       [,2]       [,3]       [,4]
[1,]  1.0000000 -0.6865422 -0.6395225  0.4010951
[2,] -0.6865422  1.0000000  0.6984153 -0.5727418
[3,] -0.6395225  0.6984153  1.0000000 -0.1538589
[4,]  0.4010951 -0.5727418 -0.1538589  1.0000000
```

```
R12
```

```
            [,1]       [,2]        [,3]      [,4]
[1,]  0.35307918 -0.6458827 -0.66378886 0.4636847
[2,] -0.06085861 -0.1140216 -0.09932185 0.1754959
```

```
E1 <-  solve(R11) %*% R12 %*% solve(R22) %*% t(R12)
E2 <-  solve(R22) %*% t(R12) %*% solve(R11) %*% R12
E1.eigen <- eigen(E1)
E2.eigen <- eigen(E2)
rho <- sqrt(E1.eigen$value[1])
rho # sample CCA
```

```
[1] 0.8142291
```

```
sqrt(E2.eigen$value[1])
```

```
[1] 0.8142291+0i
```

R 패키지 CCA의 함수 cc 를 이용하면 표본 정준상관계수와 선형변환을 위한 벡터를 구할 수 있다.

```
library(CCA)

X1 <- swiss[,c(1,6)]
X2 <- swiss[,c(2,3,4,5)]
res1 <- cc(X1,X2)
res1$cor # sample CCA
```

```
[1] 0.8142291 0.2222637
```

```
res1$xcoef # vectors for linear transformation for X1
```

```
                        [,1]        [,2]
Fertility        -0.08456464 -0.02455185
Infant.Mortality  0.05534823  0.37357101
```

```
res1$ycoef # vectors for linear transformation for X2
```

```
                   [,1]        [,2]
Agriculture  0.01985292 -0.05136137
Examination  0.02690124  0.02476760
Education    0.09414900 -0.03527373
Catholic    -0.01164052  0.01793981
```

8. 탐색적 인자 분석

```
library(tidyverse)
library(here)
library(knitr)
library(mvtnorm)
library(ggfortify)
library(HSAUR2)
library(pheatmap)
library(psych)

#아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
#font_add_google("Nanum Pen Script", "gl")
font_add_google(name = "Noto Sans KR", family = "noto")
showtext_auto()
```

인자분석(factor analysis)은 관측된 여러 변수들 간의 상관관계를 설명하기 위해, 관측된 여러 개의 변수들이 소수의 잠재적 요인(latent factors)에 의해 영향을 받는다는 가정 하에 사용되는 통계적 모형이다.

다변량 관측자료에 포함된 정보가 소수의 잠재 요인(latent factors)으로부터 생성된다고 가정하는 것은 매우 흥미로운 모형이다. 실제로 소수의 잠재적 요인이 다양한 정보들과 관련되어 있다면 현실에 나타나는 여러가지 다양한 현상들을 깊이 이해하고 잘 설명해 줄 수 있다. 인자분석은 복잡한 데이터 구조를 단순화하고 이해하는 데 도움을 준다.

또한 심리학, 사회학, 경제학에서는 자연과학과 다르게 연구의 대상이 되는 개념을 직접 측정할 수 없는 경우가 흔하게 일어난다. 예를 들어 자연과학에서 온도(temperature)라는 개념은 과학적인 정의에 따라서 다양한 방법으로 쉽고 정확하게 측정할 수 있다. 하지만 지능(intelligence), 스트레스(stress), 고통(pain), 만족도와 같은 연구 대상들은 온도와 다르게 정의도 어렵고 측정은 더 어려운 것이 현실이다.

예를 들어 학교에서 배우는 여러 과목의 시험점수들에 개인이 가지고 있는 지능(intelligence)라는 잠재변수가 공통적으로 영향을 미친다고 가정할 수 있다. 여기서 지능은 측정할 수 없으며 시험의 결과는 개인이 가진 지능에 영향을 받으며 점수로 측정값을 구할 수 있다.

이렇게 정의와 측정이 어려운 개념을 직접 관찰할 수 없는 요인(**latent factor; 인자**)으로 보고 이러한 요인에 영향을 받아 그 값을 실제로 관측할 수 있는 다양한 파생적인 변수(**manifest variables; 명시변수, 관측변수**)들을 관측할 수 있다고 가정할 수 있다. 이렇게 소수의 인자로 다수의 관측 가능한 변수들의 관계를 선형 관계로

가정하고 해석하는 분석을 인자 분석(factor analysis)라고 한다. 인자분석은 심리학, 마케팅, 경제학, 정치학 등 매우 다양한 분야에서도 중요한 분석 방법으로 사용된다.

인자 분석은 두 가지 유형으로 나뉜다. 첫 번째는 탐색적 요인분석(explanatory factor analysis)으로, 어떤 관측변수가 어떤 요인과 관련되는지에 대한 가정을 하지 않은 채 관측변수와 요인 간의 관계를 조사하는 데 사용된다. 확증적 요인 분석(confirmatory factor analysis)으로, 사전에 가정된 특정 요인 모델이 관측 변수들 간의 분산이나 상관관계에 적합한지 검증하는 데 사용된다. 이 장에서는 탐색적 요인분석만을 다루고 확증적 요인 분석은 다루지 않을 것이다.

인자분석의 목적과 방법을 간단하게 요약하면 다음과 같이 말할 수 있다.

- 가정된 인자(factor) 또는 잠재변수(latent variable)와 측정변수들의 관계를 찾는것이 탐색적 인자분석의 목적이다.
- 인자는 측정할 수 없는 변수로서 각 측정변수에 영향을 미친다고 가정한다.
- 여러 개의 관측변수에 영향을 미치는 공통의 잠재변수는 공통인자(common factor)라고 말한다.

8.1. 인자 모형

8.1.1. 단순 인자 모형

이 장에서는 단순 인자 모형(one factor model)을 사용하여 인자분석의 기본 개념을 설명한다. 단순 인자 모형은 하나의 잠재 요인(factor)이 여러 관측 변수들에 영향을 미친다고 가정하는 모델이다.

탐색적 인자 분석을 이용한 단순 인자 모형은 Spearman 이 1904년 학생들의 시험 성적과 지능에 대한 모형을 고려하면서 처음 제안되었다. 이제 Spearman 이 제안한 단순 인자 분석에 대한 통계적 가정과 모형을 살펴보자.

Spearman 은 세 과목의 시험점수에 대한 자료를 얻어서 다음과 같은 상관행렬 $\boldsymbol{R}$을 얻었다.

- X_1 : Classics
- X_2 : French
- X_3 : English

$$\boldsymbol{R} = \begin{bmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.0 \end{bmatrix}$$

지능(intelligence)을 나타내는 잠재 변수인 f가 3개의 시험 점수와 다음과 같은 관계를 가진다고 가정하는 것이 일인자 인자 모형(one factor model)이다. 모형은 다음과 같은 수식으로 표현된다.

$$\begin{aligned} X_1 &= \lambda_1 f + u_1 \\ X_2 &= \lambda_2 f + u_2 \\ X_3 &= \lambda_3 f + u_3 \end{aligned} \tag{8.1}$$

식 8.1 에서 제시된 일인자 모형에 대한 특성은 다음과 같다.

- f 는 공통 인자(common factor, latent variable)로 관측할 수 없는 확률 변수(random variable)이다.
- $\lambda_1, \lambda_2, \lambda_3$는 인자 적재값 (factor loading) 이라고 부르며 고정된 값을 가진 계수이다.
- u_i 는 임의 변동(random disturbance)으로 점수 X_i 에 반영되는 양으로 개인의 특정한 능력(specific factor) 과 단순 오차(random error)의 합으로 구성된다.

식 8.1 에서 정의된 항들의 분포 가정과 관측 변수의 분산 구조는 다음과 같다.

- X_i's 는 평균이 0 이고 분산이 1인 확률변수이다 (표준화)
- f 는 평균이 0 이고 분산이 1인 확률변수이다.
- u_i's 평균이 0 이고 분산이 ψ_i 인 확률변수이다
- u_i 들은 서로 독립이다.
- f 와 u_i 들도 서로 독립이다.

이제 관측 변수 X_i's 의 분산과 공분산을 계산하여 식 8.1 에서 제시된 모형이 상관행렬 $\boldsymbol{R}$를 어떻게 설명하는지 살펴보자.

먼저 X_i's 의 분산을 계산하면 다음과 같다.

$$\begin{aligned} V(X_i) &\equiv 1 \\ &= V(\lambda_i f + u_i) \\ &= \lambda_i^2 V(f) + V(u_i) \\ &= \lambda_i^2 + \psi_i \end{aligned}$$

다음으로 X_i's 의 공분산을 계산하면 다음과 같다.

$$\begin{aligned} Cov(X_i, X_j) &= Cov(\lambda_i f + u_i, \lambda_j f + u_j) \\ &= \lambda_i \lambda_j Cov(f, f) + \lambda_i Cov(f, u_j) + \lambda_j Cov(f, u_i) + Cov(u_i, u_j) \\ &= \lambda_i \lambda_j V(f) + 0 + 0 + 0 \\ &= \lambda_i \lambda_j \\ &\equiv corr(X_i, X_j) \quad \text{since } V(X_i) = 1 \end{aligned} \tag{8.2}$$

8.1.2. k-인자 모형

이제 식 8.1 에 나타난 일인자 모형을 확장하여 k 개의 인자를 가지는 k-인자 모형을 고려할 수 있다. 관측이 가능한 확률 변수의 개수는 q 개라고 하자.

여기서 인자 분석의 목적에 따라서 인자의 개수 k 는 가능한 확률 변수의 개수는 q 보다 작게 설정하는 것이 일반적이다 ($k \leq q$).

$$\begin{aligned} X_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + ... \lambda_{1k} f_k + u_1 \\ X_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + ... \lambda_{2k} f_k + u_2 \\ &... ... \\ X_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + ... \lambda_{qk} f_k + u_q \end{aligned} \tag{8.3}$$

식 8.3 에서 정의된 항들의 분포 가정과 관측 변수의 분산 구조는 다음과 같다.

- 확률 변수 X_i's 는 평균이 0 이고 분산이 1인 확률변수이다 (표준화)
- k-인자 벡터 $\boldsymbol{f}$ 는 평균이 $\mathbf{0}$ 이고 분산이 $\boldsymbol{I}_k$ 인 확률벡터이다.

$$E(\boldsymbol{f}) = 0, \quad Var(\boldsymbol{f}) = \boldsymbol{I}_k$$

- $\boldsymbol{u}$'s 평균이 $\mathbf{0}$ 이고 각 분산이 ψ_i 이며 서로 독립인 확률변수이다

$$E(\boldsymbol{u}) = 0, \quad Var(\boldsymbol{u}) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & ... & 0 \\ 0 & \psi_2 & ... & 0 \\ & & ... & \\ 0 & 0 & ... & \psi_q \end{bmatrix}$$

- $\boldsymbol{f}$ 와 $\boldsymbol{u}$ 들도 서로 독립이다.

식 8.3 의 k-인자 모형을 행렬식으로 표현하면 다음과 같다.

$$\boldsymbol{X} = \boldsymbol{\Lambda f} + \boldsymbol{u} \tag{8.4}$$

여기서 $\boldsymbol{\Lambda}$는 다음과 같은 $q \times k$ 행렬이다

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & ... & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & ... & \lambda_{2k} \\ & & ... & \\ \lambda_{q1} & \lambda_{q2} & ... & \lambda_{qk} \end{bmatrix}$$

식 8.3 에서 제시된 k-인자 모형과 통계적 가정에 따라서 관측 변수들의 분산 구조는 다음과 같다.

$$
\begin{aligned}
Var(X_i) &= Var(\lambda_{i1}f_1 + \lambda_{i2}f_2 + ... \lambda_{ik}f_k + u_i) \\
&= \sum_{l=1}^{q} \lambda_{il}^2 + \psi_i \\
&= h_i^2 + \psi_i \\
&\equiv 1
\end{aligned}
\tag{8.5}
$$

식 8.5 에서 나타난 분산의 두 부분을 다음과 같이 부른다.

1. $h_i^2 = \sum_{l=1}^{q} \lambda_{il}^2$를 변수의 공통성(communality)라고 부른다.

- 공통성(communality)의 의미는 관측변수들 모두가 공통으로 영향을 받는 k개의 인자들에 의해 설명되는 분산의 크기를 나타낸다. 공통성이 크다는 것은 해당 변수가 인자들에 의해 잘 설명된다는 것을 의미한다.

2. ψ_i는 임의변동(random disturbance) 또는 유일분산(unique variance)라 부른다.

- 임의변동은 변수 X_i의 분산 중에서 인자들에 의해 설명되지 않는 부분을 나타낸다. 임의변동이 크다는 것은 해당 변수가 공통 인자들이 아닌 특정한 요인에 의해 영향을 더 많이 받는다는 것을 의미한다.

각 변수간의 상관 관계는 다음과 같다. 변수간의 상관관계는 특수분산과 관계없다.

$$
\begin{aligned}
cor(X_i, X_j) &= cor(\sum_{l=1}^{q} \lambda_{il}f_l + u_i, \sum_{l=1}^{q} \lambda_{jl}f_l + u_j) \\
&= \sum_{l=1}^{q} \lambda_{il}\lambda_{jl} Var(f_l) + 0 + 0 + 0 \\
&= \sum_{l=1}^{q} \lambda_{il}\lambda_{jl}
\end{aligned}
$$

위의 결과와 관측벡터에 대한 표준화 가정을 이용하면 관측벡터 $\boldsymbol{X}$ 의 공분산은 다음과 같이 상관계수 행렬 $\boldsymbol{R}$ 로 나타난다.

$$
\begin{aligned}
\boldsymbol{\Sigma} &= Cov(\boldsymbol{X}) \\
&= Cov(\boldsymbol{\Lambda f} + \boldsymbol{u}) \\
&= \boldsymbol{\Lambda} Cov(\boldsymbol{f}) \boldsymbol{\Lambda}^t + Cov(\boldsymbol{u}) \\
&= \boldsymbol{\Lambda\Lambda}^t + \boldsymbol{\Psi} \\
&\equiv \boldsymbol{R}
\end{aligned}
\tag{8.6}
$$

여기서 $\boldsymbol{\Psi}$는 특수분산 ψ_i들을 대각원소로 가지는 대각행렬이다.

$$
\boldsymbol{\Psi} = \begin{bmatrix}
\psi_1 & 0 & ... & 0 \\
0 & \psi_2 & ... & 0 \\
 & & ... & \\
0 & 0 & ... & \psi_q
\end{bmatrix}
$$

8.1.3. 척도 불변성

지금까지 관측변수 X_i가 평균은 0 이고(언제나 관측값에서 평균을 빼면 가능하다) 분산이 1 이라고 표준화 가정을 사용하였다. 이제 관측변수 X_i가의 분산이 1이 아니라고 가정하고 X_i에 대한 척도 변환(scale transformation)을 생각해보자. 즉, 새로운 확률변수 Y_i 를 X_i에 대한 척도 변환으로 다음과 같이 정의한다.

$$Y_i = c_i X_i, \quad i = 1, 2, 3 \dots, q$$

행렬 $\boldsymbol{C}$를 c_i 들로 이루어진 대각 행렬이라고 한다면

$$\boldsymbol{C} = \begin{bmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & c_q \end{bmatrix}$$

새로운 관측값 확률벡터는 $\boldsymbol{Y} = \boldsymbol{CX}$이고 변환된 $\boldsymbol{Y}$의 공분산은 식 8.6 에 의하여 다음과 같이 쓸 수 있다.

$$\begin{aligned} Cov(\boldsymbol{Y}) &= Cov(\boldsymbol{CX}) \\ &= \boldsymbol{C}Cov(\boldsymbol{X})\boldsymbol{C}^t \\ &= \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^t\boldsymbol{C}^t + \boldsymbol{C}\boldsymbol{\Psi}\boldsymbol{C}^t \end{aligned}$$

위의 식에서 $\boldsymbol{\Lambda}_y = \boldsymbol{C}\boldsymbol{\Lambda}$로 정의하고 $\boldsymbol{u}_y = \boldsymbol{C}\boldsymbol{u}$ 로 놓으면 변환된 확률 벡터 $\boldsymbol{Y}$에 대한 인자모형은 다음과 같다.

$$\boldsymbol{Y} = \boldsymbol{\Lambda}_y \boldsymbol{f} + \boldsymbol{u}_y \tag{8.7}$$

식 8.7 을 척도 변환과 함께 풀어 쓰면 다음과 같다.

$$\begin{aligned} Y_1 &= c_1\lambda_{11}f_1 + c_1\lambda_{12}f_2 + \cdots + c_1\lambda_{1k}f_k + c_1u_1 \\ Y_2 &= c_2\lambda_{21}f_1 + c_2\lambda_{22}f_2 + \cdots + c_2\lambda_{2k}f_k + c_2u_2 \\ &\dots \\ Y_q &= c_q\lambda_{q1}f_1 + c_q\lambda_{q2}f_2 + \cdots + c_q\lambda_{qk}f_k + c_qu_q \end{aligned}$$

위의 식에서 볼 수 있듯이 척도변환을 하면 각 인자의 적재값과 임의변동의 값도 같은 척도로 변하는 것을 알 수 있다. 만약에 c_i를 관측변수의 표준편차 s_i 의 역수로 놓으면 $c_i = 1/s_i$ $\boldsymbol{Y}$의 공분산 행렬은 $\boldsymbol{X}$의 상관계수 행렬이 된다.

따라서 인자분석을 하는 경우 $\boldsymbol{X}$의 공분산 행렬을 이용하는 것과 상관계수 행렬을 이용하는 것이 일치하는 결과를 준다. 일치하는 결과라는 것은 척도변환에 의하여 언제나 대응되는 계수를 구할 수 있다는 의미이다 (**척도의 불변성, scale invariance**)

참고로 주성분 분석은 공분산 행렬을 이용하는 경우와 상관계수 행렬을 이용하는 경우 척도가 달라지기 때문에 동일한 결과를 얻을 수 없다.

8.1.4. 인자의 비유일성

임의의 $k \times k$의 직교행렬 $\boldsymbol{P}$을 생각하고 인자모형을 변환하여 보자

$$\begin{aligned}\boldsymbol{X} &= \boldsymbol{\Lambda f} + \boldsymbol{u} \\ &= \boldsymbol{\Lambda P P}^t \boldsymbol{f} + \boldsymbol{u} \\ &= (\boldsymbol{\Lambda P})(\boldsymbol{P}^t \boldsymbol{f}) + \boldsymbol{u} \\ &= \boldsymbol{\Lambda}_1 \boldsymbol{f}_1 + \boldsymbol{u}\end{aligned}$$

여기서 $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda P}$이고 $\boldsymbol{f}_1 = \boldsymbol{P}^t \boldsymbol{f}$이다. 위의 새로운 인자 모형에서의 인자 $\boldsymbol{f}_1$의 분포와 관측변수 $\boldsymbol{X}$의 공분산은 원래 인자 $\boldsymbol{f}$의 분포와 동일한다.

$$E(\boldsymbol{f}_1) = E(\boldsymbol{P}^t \boldsymbol{f}) = \boldsymbol{P}^t E(\boldsymbol{f}) = 0, \quad Cov(\boldsymbol{f}_1) = \boldsymbol{P}^t Var(\boldsymbol{f}) \boldsymbol{P} = \boldsymbol{P}^t \boldsymbol{P} = \boldsymbol{I}$$

따라서 새로운 인자 모형에서 관측벡터 $\boldsymbol{X}$의 공분산은 다음과 같다.

$$Cov(\boldsymbol{X}) = Cov(\boldsymbol{\Lambda}_1 \boldsymbol{f}_1 + \boldsymbol{u}) = Cov(\boldsymbol{\Lambda f} + \boldsymbol{u}) = \boldsymbol{\Lambda \Lambda}^t + \boldsymbol{\Psi}$$

따라서 인자적재값은 같은 자료라도 유일하게 존재하지 않는다(non-uniquesness). 더나아가 인자의 적재행렬에 제한 조건을 주면 유일하게 존재할 수 있다(인자의 회전; factor rotation)

8.2. 모형의 추정

이제 인자 모형을 추정하는 방법에 대해서 알아보자.

8.2.1. 단순 인자모형

먼저 식 8.1 에 나타난 단순 인자 모형과 공분산에 대한 결과 식 8.2 를 고려하면 다음과 같은 방정식을 유도할 수 있다.

$$
\begin{aligned}
\boldsymbol{R} &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^t + \boldsymbol{\Psi} \\
&= \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1^2 + \psi_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 \\ \lambda_2\lambda_1 & \lambda_2^2 + \psi_2 & \lambda_2\lambda_3 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + \psi_3 \end{bmatrix}
\end{aligned}
$$

실제 관측한 3개의 시험점수의 상관행렬 $\boldsymbol{R}$를 위의 식에 대입하면 다음과 같은 방정식을 얻는다.

$$
\begin{aligned}
\lambda_1\lambda_2 &= 0.83 \\
\lambda_1\lambda_3 &= 0.78 \\
\lambda_2\lambda_3 &= 0.67 \\
\lambda_1^2 + \psi_1 &= 1 \\
\lambda_2^2 + \psi_2 &= 1 \\
\lambda_3^2 + \psi_3 &= 1
\end{aligned}
$$

위의 비선형 방정식 시스템을 풀면 다음과 같은 해를 얻는다. 이러한 해로 식 8.1 에 나타난 단순 인자 모형의 인자와 분산성분을 추정할 수 있다.

$$
\begin{aligned}
\hat{\lambda}_1 &= 0.99, \quad \hat{\psi}_1 = 0.02 \\
\hat{\lambda}_2 &= 0.84, \quad \hat{\psi}_2 = 0.3 \\
\hat{\lambda}_3 &= 0.79, \quad \hat{\psi}_3 = 0.38
\end{aligned}
$$

8.2.2. 최대 가능도 추정법

탐색적 인자분석에서 인자 적재값과 특수분산을 추정하는 방법으로 최대 가능도 추정법(Maximum Likelihood Estimation; MLE)을 사용할 수 있다. 최대 가능도 추정법은 관측된 데이터가 주어진 모형에 의해 생성될 확률을 최대화하는 모수 값을 찾는 방법이다.

다변량 정규분포 모형에서 분산 행렬 $\boldsymbol{\Sigma}$가 인자 모형에 의해 식 8.6 과 같이 나타난다고 가정하고 최대 가능도 추정법을 적용할 수 있다.

이 장에서는 최대 가능도 추정법에 대한 자세한 설명은 생략하고 주성분 분석을 이용한 간단한 모형 추정법을 다음 절에 소개하려고 한다.

8.2.3. 주성분 인자분석

탐색적 인자분석에서 요인의 초기 추정값을 얻는 방법으로 주성분분석을 사용할 수 있다. 주성분분석를 통해 얻은 상위 몇 개의 주성분이 데이터의 분산을 대부분 설명한다면, 이 주성분들을 잠재요인의 초기 근사치로 사용할 수 있다.

이러한 접근 방법을 주성분 인자분석(Principal Component Factor Analysis) 이라고 부른다.

q-차원의 임의 벡터 $\boldsymbol{X}^t = (X_1, X_2, \dots, X_q)$ 가 평균이 0이고 공분산이 $\boldsymbol{\Sigma}$이라 가정하자.

먼저 표본 공분산 행렬 $\hat{\boldsymbol{\Sigma}}$을 이용하여 주성분 분석을 실시한다.

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{12}X_2 + \dots a_{1q}X_q \\ Z_2 &= a_{21}X_1 + a_{22}X_2 + \dots a_{2q}X_q \\ &\dots \\ Z_q &= a_{q1}X_1 + a_{q2}X_2 + \dots a_{qq}X_q \end{aligned}$$

위의 주성분 분석의 결과를 행렬식으로 나타내면 다음과 같다.

$$\boldsymbol{Z} = \boldsymbol{AX}$$

행렬 $\boldsymbol{A}$의 역행렬은 Section 6.2.4 에 의하여 표본 공분산 행렬에 대한 고유벡터 행렬이므로 직교행렬이다 ($\boldsymbol{AA}^t = \boldsymbol{I}$). 따라서 다음과 같은 역변환을 고려한다.

$$\boldsymbol{X} = \boldsymbol{A}^t\boldsymbol{Z}$$

위의 식을 다시 쓰면 다음과 같다.

$$\begin{aligned} X_1 &= a_{11}Z_1 + a_{21}Z_2 + \dots a_{q1}Z_q \\ X_2 &= a_{12}Z_1 + a_{22}Z_2 + \dots a_{q2}Z_q \\ &\dots \\ X_q &= a_{1q}Z_1 + a_{2q}Z_2 + \dots a_{qq}Z_q \end{aligned}$$

이제 q 개의 주성분들 중에서 $k < q$개의 주성분을 선택한다. 선택된 주성분을 인자로 생각할 수 있다.

$$\begin{aligned} X_1 &= a_{11}Z_1 + a_{21}Z_2 + \dots a_{k1}Z_k + u_1 \\ X_2 &= a_{12}Z_1 + a_{22}Z_2 + \dots a_{k2}Z_k + u_2 \\ &\dots \\ X_q &= a_{1q}Z_1 + a_{2q}Z_2 + \dots a_{kq}Z_k + u_3 \end{aligned} \tag{8.8}$$

여기서 유의할 점은 u_i 는 나머지 주성분 $Z_{k+1}, Z_{k+2}, \dots, Z_q$의 선형조합으로 사실 서로 독립은 아니다.

마지막으로 인자의 분포 조건에 만추기 위하여 Z_i들을 표준화한다. i 번째 주상분의 분산은 표본 공분산 행렬의 i 번째 고유값이므로 ($Var(Z_i) = \lambda_i$) 각 Z_i를 고유값의 제곱근으로 나누어 주고 각 계수에 곱해준다.

$$f_i = Z_i/\sqrt{\lambda_i}, \quad \lambda_{ij} = \sqrt{\lambda_i}a_{ji}$$

이렇게 표준화하면 다음과 같이 주성분을 이용한 인자모형을 구할 수 있다

$$
\begin{aligned}
X_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + ... \lambda_{1k} f_k + u_1 \\
X_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + ... \lambda_{2k} f_k + u_2 \\
&... \\
X_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + ... \lambda_{qk} f_k + u_q
\end{aligned}
$$

8.2.4. 인자의 선택과 회전

8.2.4.1. 인자분석 개수의 추정

관측된 공분산이나 상관관계를 적절히 표현할 수 있는 인자의 개수 k의 결정이 중요하다. 인자의 개수를 결정하는 방법으로 다음과 같은 방법들이 있다.

먼저, 최대가능도 추정법에서 가능도비 검정을 이용하여 인자의 개수를 가설 검정의 형식으로 정할 수 있다. 귀무가설 H_0 에서 인자의 개수를 k개로 하여 가능도비 검정이 가능하다. 없다.

두 번째로 탐색적 인자 분석에서는 개수에 대한 여러 개의 후보값을 고려하여 주성분 분서과 유사한 방법으로 적절한 개수를 결정한다.

8.2.4.2. 인자의 회전

Section 8.1.4 에서 설명한 바와 같이 인자 모형에서 인자의 적재값은 유일하지 않다. 인자의 적재값을 유일하게 만들기 위하여 여러 가지 제약식을 사용할 수 있다.

식 8.4 에서 다음과 같이 주어지는 행렬 G이 **대각행렬이 되고 대각원소들이 작아지는 순서로** 되는 인자를 선택하면 주성분분석에서와 유사하게 인자의 중요성이 분산을 설명하는 순서대로 정렬된다.

$$
\boldsymbol{G} = \boldsymbol{\Lambda}^t \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}, \quad \boldsymbol{G} \text{ is diagonal with decreasing elements}
$$

이러한 제약 조건은 첫 번째 인자가 관측 변수들의 공분산에 최대 기여를 하도록 설정하며, 두 번째 인자는 첫 번째 인자와 상관관계가 없으면서 이 분산에 최대 기여를 하도록 한다. 이와 같은 과정이 반복된다(이러한 조건은 주성분분석과 내우 유사하다). 위의 제약조건은 Λ가 열의 부호 변경 가능성을 제외하고 유일하게 결정되도록 보장한다.

참고할 사항은 위와 같은 제약 조건은 참모형, 즉 유일한 $\boldsymbol{\Lambda}$ 을 찾는 방법이 아니라 해를 유일하게 만들어 주는 임의의 제약 조건(arbitrary constraint) 를 적용하는 것이다.

인자의 적재값이 유일하게 주어지는 제약조건은 다음과 같이 인자 적재값을 구성하는 행렬 $\boldsymbol{\Lambda}$에 직교행렬을 곱해주는 방식으로 구현될 수 있으며 이를 **인자의 회전(factor rotation)** 이라고 한다.

$$
\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda} \boldsymbol{P}
$$

인자의 회전은 모형의 계수에 대한 근본적인 특성을 변경하지 않으면서 해석 가능성을 높이는 과정이다. 각 변수가 하나의 요인에 높은 적재를 가지며, 요인 적재값이 상재적으로 크거나 또는 거의 0에 가까워서 중간값이 거의 없을

때 해석이 더 직관적이다. 이러한 해석에서의 용이성에 대한 관점은 주성분 분석에서 성분의 의미를 해석하는 점과 매우 유사하다.

결론적으로 인자의 회전은 인자의 적재값들이 **모형의 해석을 가능한 한 단순할 수 있는 방향**으로 선택하는 것이 바람직하다. 다음과 같은 성질을 구현할 수 있는 회전이 있다면 해석이 용이할 것이다.

- 각 행 또는 인자적재행렬은 적어도 하나의 0을 포함
- 인자적재행렬의 각 열은 적어도 k개의 0을 포함
- 인자적재행렬의 열들을 비교할 때 변수등들이 대비되면 좋다(예: 한 열에는 크고 다른 열에는 작은 값)

인자 회전은 다양하고 많은 기준과 방법이 있지만 대표적인 방법은 직교회전(orthogonal rotation)으로 회전된 인자들이 서로 상관되지 않게 만드는 방법이다.또한 사각회전(oblique rotation)도 가능하며 이는 상관된 인자를 허용한다.

직교회전은 원래 구한 적재행렬을 직교 회전하면 얻을 수 있다. 직교회전의 대표적인 방법은 Varimax 회전으로 적재값이 가능한 한 큰값을 가지거나 또는 0에 가깝운 값을 가지게 하여 해석이 쉽게되도록 하는 목적을 가지고 있다.

인자의 회전을 사용하는 예제로 인공적인 자료를 만들어서 회전 전과 후의 적재값을 비교해보자.

대학생의 알코올/담배/대마/LSD/코카인에 대한 사용 정도를 4단계(0=never, 3=regularly)의 정도와 5문항으로 구성된 합성 자료를 만들어 보자. 이 경우 2개의 인자를 가지는 모형 구조(일상적 vs 비교적 위험)로부터 데이터를 생성하고, 상관 행렬을 이용한 최대가능도 추정에서 인자를 회전하지 않는 경우와 varimax 회전을 적용한 차이를 보여주려고 한다.

$$X_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + u_i, \quad i = 1, 2, 3, 4, 5$$

위의 모형에 대하여 다음과 같은 분포와 가정을 사용한다.

$$\boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix} \right), \quad \boldsymbol{\Lambda} = \begin{bmatrix} 0.20 & 0.70 \\ 0.25 & 0.75 \\ 0.70 & 0.30 \\ 0.80 & 0.10 \\ 0.85 & 0.10 \end{bmatrix}$$

다음은 자료를 생성하는 R 코드이다.

```
# 패키지
library(psych)

# 생성된 자료의 재현성을 위하여 난수 시드 고정
set.seed(2025)

# 표본 크기
```

```r
n <- 400

#
# 잠재 요인 2개: F1=위험·환각성, F2=일상(음주/흡연)
SigmaF <- matrix(c(1, 0.25,
                   0.25, 1), 2, 2)    # 요인 간 약한 양의 상관(필요시 0으로 직교 가정)

# 잠재 요인 점수 생성
F <- MASS::mvrnorm(n, mu = c(0,0), Sigma = SigmaF)

# 인자 적제값 매트릭스(5문항)
#   F1(loadings): marijuana, lsd, cocaine
#   F2(loadings): alcohol, cigarette
L <- matrix(c(
  # F1   F2
   0.20, 0.70,   # alcohol
   0.25, 0.75,   # cigarette
   0.70, 0.30,   # marijuana
   0.80, 0.10,   # lsd
   0.85, 0.10    # cocaine
), nrow=5, byrow=TRUE)

# 관측 변수 = L %*% F + 오차
err <- MASS::mvrnorm(n, mu = rep(0,5), Sigma = diag(5)*0.5)
Z <- F %*% t(L)          # n x 5
Z <- scale(Z) + scale(err)
colnames(Z) <- c("alcohol","cigarette","marijuana","lsd","cocaine")

# 연속 -> 4점 서열화(0,1,2,3)
cuts <- c(-Inf, -0.2, 0.6, 1.4, Inf)
X <- apply(Z, 2, function(v) as.integer(cut(v, breaks = cuts, labels = FALSE)) - 1)
X <- as.data.frame(X)

head(X)
```

```
  alcohol cigarette marijuana lsd cocaine
1       2         3         2   2       1
2       2         1         1   1       1
3       1         0         2   3       3
4       1         3         1   2       2
5       1         2         1   1       1
6       1         0         0   0       0
```

이제 생성된 자료에 대하여 인자분석을 수행해보자. 먼저 상관행렬을 계산하고 psych 패키지에 있는 fa 함수를 이용하여 최대가능도 추정법을 이용한 인자분석을 수행한다.

```
#  상관 계수 행렬
R <- cor(X)
R
```

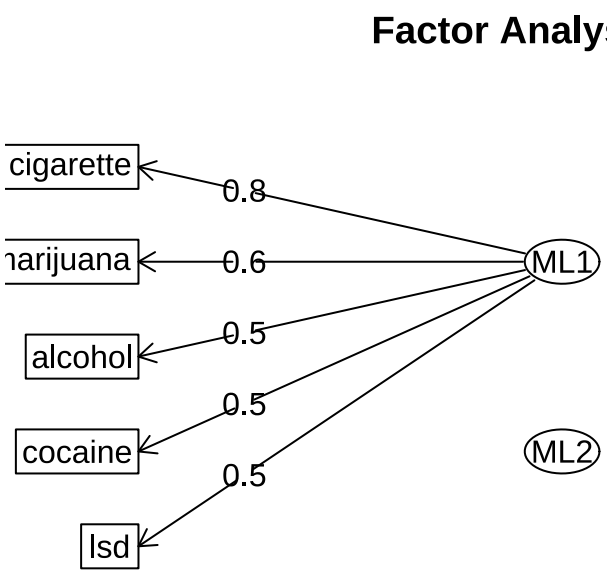
```
            alcohol cigarette marijuana       lsd   cocaine
alcohol   1.0000000 0.4362218 0.3260260 0.3131031 0.3216480
cigarette 0.4362218 1.0000000 0.3764078 0.3001491 0.2764881
marijuana 0.3260260 0.3764078 1.0000000 0.3942864 0.4267750
lsd       0.3131031 0.3001491 0.3942864 1.0000000 0.4648938
cocaine   0.3216480 0.2764881 0.4267750 0.4648938 1.0000000
```

```
# ML 요인분석 + 2개의 요인 + 무회전
fa_fit1 <- fa(R, nfactors = 2, fm = "ml", rotate = "none")
print(fa_fit1, digits = 2, cut = 0.30)
```

```
Factor Analysis using method =  ml
Call: fa(r = R, nfactors = 2, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
           ML1   ML2   h2   u2 com
alcohol   0.55       0.31 0.69 1.0
cigarette 0.84       0.79 0.21 1.2
marijuana 0.55       0.39 0.61 1.5
lsd       0.50  0.41 0.42 0.58 1.9
cocaine   0.51  0.51 0.52 0.48 2.0

                       ML1  ML2
SS loadings           1.82 0.60
Proportion Var        0.36 0.12
Cumulative Var        0.36 0.49
Proportion Explained  0.75 0.25
Cumulative Proportion 0.75 1.00

Mean item complexity =  1.5
Test of the hypothesis that 2 factors are sufficient.

df null model =  10  with the objective function =  0.99
df of  the model are 1  and the objective function was  0

The root mean square of the residuals (RMSR) is  0
```

```
The df corrected root mean square of the residuals is  0

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 ML1  ML2
Correlation of (regression) scores with factors   0.92 0.76
Multiple R square of scores with factors          0.84 0.58
Minimum correlation of possible factor scores     0.68 0.17
```

```
# 시각화
fa.diagram(fa_fit1)
```



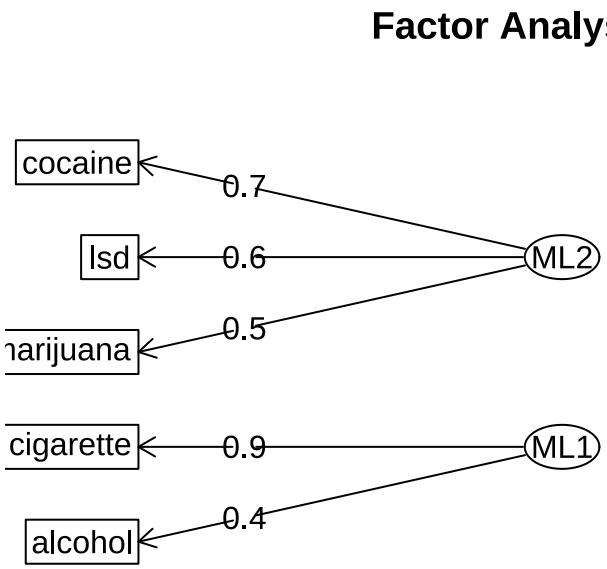
위의 결과를 보면 인자의 적재값이 모든 변수에 걸쳐 분산되어 있어 해석이 어렵다. 이제 varimax 회전을 적용하여 인자적재값을 다시 계산해보자.

```
# ML 요인분석 + 2개의 요인 + varimax 회전
fa_fit2 <- fa(R, nfactors = 2, fm = "ml", rotate = "varimax")
print(fa_fit2, digits = 2, cut = 0.30)
```

```
Factor Analysis using method =  ml
Call: fa(r = R, nfactors = 2, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
           ML2  ML1   h2   u2 com
alcohol   0.36 0.42 0.31 0.69 2.0
cigarette      0.87 0.79 0.21 1.1
marijuana 0.54 0.31 0.39 0.61 1.6
```

```
lsd       0.62      0.42 0.58 1.2
cocaine   0.70      0.52 0.48 1.1

                       ML2  ML1
SS loadings           1.33 1.10
Proportion Var        0.27 0.22
Cumulative Var        0.27 0.49
Proportion Explained  0.55 0.45
Cumulative Proportion 0.55 1.00

Mean item complexity =  1.4
Test of the hypothesis that 2 factors are sufficient.

df null model =  10  with the objective function =  0.99
df of  the model are 1  and the objective function was  0

The root mean square of the residuals (RMSR) is  0
The df corrected root mean square of the residuals is  0

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                 ML2  ML1
Correlation of (regression) scores with factors   0.81 0.88
Multiple R square of scores with factors          0.66 0.77
Minimum correlation of possible factor scores     0.31 0.54
```

```r
# 시각화
fa.diagram(fa_fit2)
```



위의 결과에서 볼 수 있듯이 인자의 회전을 적용한 후에는 각 인자에 대한 해석이 훨씬 용이해진 것을 알 수 있다. 첫 번째 인자는 대마초, LSD, 코카인과 강하게 관련되어 위험·환각성과 관련된 인자로 해석할 수 있고, 두 번째 인자는 알코올과 담배와 강하게 관련되어 일상적인 물질 사용과 관련된 인자로 해석할 수 있다.

-

8.2.4.3. 인자점수의 추정

인자 회전을 한 뒤에 인자의 적재값과 인자의 예측값($\hat{f}$)을 이용하여 인자점수(factor score)를 계산할 수 있다. 이러한 인자 점수는 주성분 분석의 점수와 유사하게 사용된다.

$$\hat{X} = \Lambda \hat{f}$$

References

A. 행렬의 기초

이 장에서는 회귀분석의 이론 전개에 필요한 행렬 이론과 선형 대수의 기초에 대하여 알아볼 것이다.

A.1. 벡터와 행렬

다음 p-차원 벡터(vector) 또는 열벡터(column vector) $\boldsymbol{a}$ 는 p개의 원소 $a_1, a_2, \dots, a_p$ 를 하나의 열(column)에 배치한 형태를 가진 개체이다.

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{A.1}$$

차원이 $n \times p$ 인 행렬 $\boldsymbol{A}$ 는 다음과 같이 n개의 행과 p 개의 열에 원소 a_{ij}를 다음과 같이 배치한 형태를 가진다.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

A.2. 두 행렬의 덧셈

두 행렬 $\boldsymbol{A}$ 와 $\boldsymbol{B}$ 를 더하는 규칙은 다음과 같다.

- 두 행렬 $\boldsymbol{A}$ 와 $\boldsymbol{B}$ 는 행과 열의 갯수가 같아야 한다.
- $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C}$ 라고 하면, 덧셈의 결과로 만들어진 행렬 $\boldsymbol{C}$는 두 행렬과 같은 수의 행과 열을 가지면 각 원소는 다음과 같다.

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C} \quad \rightarrow \quad c_{ij} = a_{ij} + b_{ij}$$

A.3. 스칼라곱

임의의 실수 λ (스칼라)가 주어졌을 때, λ 와 행렬 $\boldsymbol{A}$의 스칼라곱(scalar product) 는 행렬의 모든 원소에 λ 를 곱해준 행렬로 정의된다.

예를 들어 $\lambda = 2$, $\boldsymbol{A} \in \mathbb{R}^{2\times 3}$ 인 경우

$$\lambda \boldsymbol{A} = 2 \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ -2 & 0 & 4 \end{bmatrix}$$

A.4. 벡터와 행렬의 곱셈

$n \times p$ 인 행렬 $\boldsymbol{A}$ 와 p-차원 벡터(vector) $\boldsymbol{b}$는 다음과 같이 두 개의 서로 다른 형태로 나타낼 수 있다.

A.4.1. 행과 열의 내적

먼저 행렬과 벡터의 곱셈은 행렬 $\boldsymbol{A}$ 의 행벡터와 벡터 $\boldsymbol{b}$ 의 내적(inner product)로 나타낼 수 있다.

$$\begin{aligned}
\boldsymbol{Ab} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{r}_1^t \\ \boldsymbol{r}_2^t \\ \vdots \\ \boldsymbol{r}_n^t \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \qquad \text{where } \boldsymbol{r}_i^t = \begin{bmatrix} a_{i1} & a_{i2} & \dots & a_{ip} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{r}_1^t \boldsymbol{b} \\ \boldsymbol{r}_2^t \boldsymbol{b} \\ \vdots \\ \boldsymbol{r}_n^t \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^p a_{1j} b_j \\ \sum_{j=1}^p a_{2j} b_j \\ \vdots \\ \sum_{j=1}^p a_{nj} b_j \end{bmatrix} \\
&= \begin{bmatrix} < \boldsymbol{r}_1, \boldsymbol{b} > \\ < \boldsymbol{r}_2, \boldsymbol{b} > \\ \vdots \\ < \boldsymbol{r}_n, \boldsymbol{b} > \end{bmatrix}
\end{aligned}$$

위에서 $< \boldsymbol{a}, \boldsymbol{b} >$ 는 다음과 같은 두 벡터의 내적(inner product)을 의미한다.

$$< \boldsymbol{a}, \boldsymbol{b} > = \boldsymbol{a}^t \boldsymbol{b} = \sum_{i=1}^p a_i b_i$$

A.4.2. 열벡터의 선형조합

이제 행렬과 벡터의 곱셈을 행렬을 구성하는 열벡터들의 선형조합(linear combination)으로 나타낼 수 있다.

$$
\begin{aligned}
\boldsymbol{Ab} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 & \dots & \boldsymbol{c}_p \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \quad \text{where } \boldsymbol{c}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \\
&= b_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + b_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \cdots + b_p \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{np} \end{bmatrix} \\
&= b_1 \boldsymbol{c}_1 + b_2 \boldsymbol{c}_2 + \cdots + b_p \boldsymbol{c}_p
\end{aligned}
$$

A.5. 행렬의 전치

$\boldsymbol{A}^t$는 행렬의 전치(transpose)를 나타낸다. 행렬의 전치는 원소의 행과 열을 바꾸어 만든 행렬이다.

$$
\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \dots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix} = \{a_{ij}\}_{n \times p} \quad \rightarrow \quad \boldsymbol{A}^t = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & & \dots \\ a_{1p} & a_{2p} & \dots & a_{np} \end{bmatrix} = \{a_{ji}\}_{p \times n}
$$

A.6. 행렬의 곱셈

먼저 두 행렬 $\boldsymbol{A}$ 와 $\boldsymbol{B}$ 의 곱셈

$$
\boldsymbol{A} \times \boldsymbol{B} \equiv \boldsymbol{AB}
$$

을 정의하려면 다음과 같은 조건이 만족되어야 한다.

- 행렬 $\boldsymbol{A}$ 의 열의 갯수와 행렬 $\boldsymbol{B}$ 의 행의 갯수가 같아야 한다

따라서 두 행렬의 곱셈은 순서를 바꾸면 정의 자체가 안될 수 있다.

이제 두 행렬 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ 와 $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ 의 곱셈은 다음과 같이 정의된다.

$$\boldsymbol{AB} = \boldsymbol{C}$$

행렬 $\boldsymbol{C}$ 는 m 개의 행과 k개의 열로 구성된 행렬이며($\boldsymbol{C} \in \mathbb{R}^{m \times k}$) 각 원소 c_{ij}는 다음과 같이 정의된다.

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lk}, \quad i = 1, 2, \dots, m; j = 1, 2, \dots, k$$

먼저 간단한 예제로 다음과 같은 두 개의 행렬의 곱을 생각해 보자.

$$\boldsymbol{AB} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} (1)(0) + (2)(-1) & (1)(1) + (2)(2) \\ (3)(0) + (4)(-1) & (3)(1) + (4)(2) \end{bmatrix} = \begin{bmatrix} -2 & 5 \\ -4 & 11 \end{bmatrix}$$

곱하는 순서를 바꾸어 계산해 보자.

$$\boldsymbol{BA} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} (0)(1) + (1)(3) & (0)(2) + (1)(4) \\ (-1)(1) + (2)(3) & (-1)(2) + (2)(4) \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix}$$

위 두 결과를 보면 행렬의 곱셈에서는 교환법칙이 성립하지 않음을 알 수 있다.

이제 차원이 다른 두 행렬의 곱셈을 살펴보자.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}$$

두 행렬의 곱셈은 다음과 같이 계산할 수 있다.

$$\boldsymbol{AB} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 5 \end{bmatrix}$$

두 행렬의 곱하는 순서를 바꾸면 차원이 전혀 다른 행렬이 얻어진다.

$$\boldsymbol{BA} = \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 4 & 2 \\ -2 & 0 & 2 \\ 3 & 2 & 1 \end{bmatrix}$$

행렬의 곱셈은 교환법칙이 성립하지 않는다.

$$\boldsymbol{AB} \neq \boldsymbol{BA} \tag{A.2}$$

Caution

교환법칙이 성립하지 않는다는 의미는 식 A.2 이 언제나 성립한다는 의미는 아니다. 아래와 같이 특별한 경우 교환법칙이 성립하는 경우도 있다.

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$$

- 행렬의 곱셈은 결합법칙과 배분법칙은 성립한다.

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$$

$$(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$$

A.7. 단위벡터와 항등행렬

i번째 단위벡터 $\boldsymbol{e}_i$를 정의하자. 단위벡터 $\boldsymbol{e}_i$는 n- 차원 벡터로서 i번째 원소만 1이고 나머지는 0인 벡터이다.

$$\boldsymbol{e}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

즉 n-차원 항등행렬 $\boldsymbol{I}$는 n개의 단위벡터들을 모아놓은 것이다. 단위행렬은 대각원소가 1이고 나머지는 0인 정방행렬이다.

$$\boldsymbol{I} = [\boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \ldots \;\; \boldsymbol{e}_n]$$

A.8. 대각합

$\boldsymbol{A} = \{a_{ij}\}$를 $n \times n$ 정방행렬(square matrix)인 경우, 행렬의 대각 원소(diagonal element)들의 합(trace)을 $tr(\boldsymbol{A})$로 표시한다.

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$$

두 행렬의 덧셈(뺄셈)에 대한 대각합에 대한 성질들은 다음과 같다.

$$tr(\boldsymbol{A} \pm \boldsymbol{B}) = tr(\boldsymbol{A}) \pm tr(\boldsymbol{B})$$

> **Caution**
>
> 행렬의 곱셈은 일반적으로 교환법칙이 성립하지 않지만 대각합의 연산은 교환법칙이 성립한다.
>
> $$tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$$

대각합은 교환법칙이 성립히기 때문에 다음과 같은 성질이 성립한다.

$$\mathsf{tr}(\boldsymbol{AKL}) = \mathsf{tr}(\boldsymbol{KLA})$$

벡터의 연산에서도 대각합의 교환법칙이 성립되어 다음과 같은 유용한 식이 성립한다.

$$\mathsf{tr}\left(\boldsymbol{x}\boldsymbol{y}^t\right) = \mathsf{tr}\left(\boldsymbol{y}^t\boldsymbol{x}\right) = \boldsymbol{y}^t\boldsymbol{x} \in \mathbb{R}.$$

대각합의 교환법칙때문에 어떤 행렬의 앞에 특정 행렬을 곱하고, 뒤에 역행렬을 곱해도 대각합은 변하지 않는다.

$$\mathsf{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{AS}\right) = \mathsf{tr}\left(\boldsymbol{ASS}^{-1}\right) = \mathsf{tr}(\boldsymbol{A})$$

대각합에 대한 그 밖의 성질들은 다음과 같다.

- $\mathsf{tr}(\alpha\boldsymbol{A}) = \alpha\,\mathsf{tr}(\boldsymbol{A}), \alpha \in \mathbb{R}$ for $\boldsymbol{A} \in \mathbb{R}^{n\times n}$
- $\mathsf{tr}\left(\boldsymbol{I}_n\right) = n$

A.9. 행렬식

$\boldsymbol{A}$의 행렬식(determinant)을 $det(\boldsymbol{A}) = |\boldsymbol{A}|$로 표기한다.

이차원 행렬 $\boldsymbol{A}$ 의 행렬식은 다음과 같이 계산한다.

$$\mathsf{det}(\boldsymbol{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

만약 행렬 $\boldsymbol{A}$가 대각행렬(diagonal matrix)이면 $|\boldsymbol{A}|$는 행렬의 대각원소의 곱이다 ($|\boldsymbol{A}| = \prod a_{ii}$).

두 행렬의 곱의 행렬식은 각 행렬의 행렬식의 곱이다.

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$$

행렬식에 대한 유용한 공식들은 다음과 같다.

- $|\boldsymbol{A}^t| = |\boldsymbol{A}|$
- $|c\boldsymbol{A}| = c^n|\boldsymbol{A}|$

만약 행렬 $\boldsymbol{A}$가 다음과 같은 분할행렬(partitioned matrix) 의 형태를 가지면

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{bmatrix}$$

행렬 $\boldsymbol{A}$의 행렬식은 다음과 같이 주어진다.

$$|\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22}|$$

다음과 같은 행렬식에 대한 공식도 유용하다. p-차원 행렬 $\boldsymbol{A}$가 역행렬이 존재하고 벡터 $\boldsymbol{u}$ 와 $\boldsymbol{v}$ 에 대하여

$$|\boldsymbol{A} + \boldsymbol{uv}^t| = |\boldsymbol{A}|(1 + \boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{u})$$

A.10. 역행렬

만약 정방행렬 $\boldsymbol{A}$가 다음과 같은 조건을 만족하면 정칙행렬(invertible matrix 또는 nonsingular matrix)이라고 부른다.

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

이경우 $\boldsymbol{A}^{-1}$ 를 행렬 $\boldsymbol{A}$의 역행렬(inverse matrix) 이라고 정의한다.

역행렬이 존재할 조건은 행렬 $\boldsymbol{A}$의 행렬식이 0이 아니어야 한다.

A.11. 직교행렬

만약 정방행렬 $\boldsymbol{P}$가 다음과 같은 조건을 만족하면 직교행렬(orthogonal matrix)라고 부른다.

$$\boldsymbol{P}\boldsymbol{P}^t = \boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$$

직교행렬의 정의에서 주의할 점은 $\mathbf{P}\ \mathbf{P}\hat{}\mathrm{t} = \mathbf{I}\ \$$ 와 $\boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$ 이 모두 성립하해야 한다는 점이다.

행렬 $\boldsymbol{P}$ 의 역행렬은 $\boldsymbol{P}^t$ 이다.

$$\boldsymbol{P}^{-1} = \boldsymbol{P}^t$$

만약 $\boldsymbol{P}$가 직교행렬이면 다음과 같은 성질을 가진다.

- $|\boldsymbol{P}| = \pm 1$, 왜냐하면
$$|\boldsymbol{P}\boldsymbol{P}^t| = |\boldsymbol{P}||\boldsymbol{P}^t| = |\boldsymbol{P}|^2 = |\boldsymbol{I}| = 1$$
- 임의의 정방행렬 $\boldsymbol{A}$에 대하여 다음이 성립한다.
$$tr(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^t) = tr(\boldsymbol{A}\boldsymbol{P}^t\boldsymbol{P}) = tr(\boldsymbol{A})$$

A.12. 벡터의 선형독립

n 개의 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 의 선형결합(또는 선형결합, linear combination)이란 각 벡터에 스칼라를 곱하여 더한 것들이다.

만약 $r_1, r_2, \ldots, r_n$ 가 임의의 실수일 때, 다음과 같은 형태의 선형식을 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$의 선형결합(linear combination)이라고 한다:

$$r_1\boldsymbol{v}_1 + r_2\boldsymbol{v}_2 + \cdots + r_n\boldsymbol{v}_n \tag{A.3}$$

Definition A.1 (벡터의 선형독립과 선형종속)**.** n 개의 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 가 있다고 하자. 만약 다음 식이 만약 모두 0인 n개의 스칼라 $r_1, r_2, \ldots, r_n$에 대해서만 성립하면 n개 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 들은 선형독립(linearly independent)라고 한다.

$$r_1\boldsymbol{v}_1 + r_2\boldsymbol{v}_2 + \cdots + r_n\boldsymbol{v}_n = \mathbf{0} \quad \Longleftrightarrow r_1 = r_2 = \cdots = r_n = 0 \tag{A.4}$$

또한 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 가 선형독립이 아니면 선형종속(linear dependent)라고 한다. 벡터 $\boldsymbol{v}_1,\ \boldsymbol{v}_2,\ \ldots\ ,\boldsymbol{v}_n$ 가 선형종속이면 **모두 0이 아닌** $r_1, r_2, \ldots, r_n$ 이 존재하여 다음이 성립한다는 것이다.

$$\exists\, r_1, r_2, \ldots, r_n \in R \quad \text{s.t. } (r_1, r_2, \ldots, r_n) \neq \mathbf{0}, r_1\boldsymbol{v}_1 + r_2\boldsymbol{v}_2 + \cdots + r_n\boldsymbol{v}_n = \mathbf{0} \tag{A.5}$$

■

예를 들어 다음과 같이 주어진 3개의 3-차원 벡터들은 선형종속이다.

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix}1\\0\\1\end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix}3\\2\\5\end{bmatrix} \tag{A.6}$$

왜냐하면 다음과 같이 모두 0이 아닌 스칼라에 의해서 다음 식이 성립하기 때문이다. 즉 벡터 $\boldsymbol{v}_3$는 $\boldsymbol{v}_2$ 에 2를 곱하여 $\boldsymbol{v}_1$에 더한 값과 같다.

$$\boldsymbol{v}_3 = \boldsymbol{v}_1 + 2\boldsymbol{v}_2 \quad \Longleftrightarrow \quad \boldsymbol{v}_1 + 2\boldsymbol{v}_2 - \boldsymbol{v}_3 = 0$$

이제 다음과 같이 주어진 3개의 3-차원 벡터들은 선형독립이다. 즉 3개 벡터의 선형 조합이 0이 될 수 있도록 만드는 스칼라는 모두 0인 경우 밖에 없다.

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix}1\\0\\1\end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix}3\\2\\4\end{bmatrix} \tag{A.7}$$

이제 다음과 같이 주어진 4개의 3-차원 벡터들은 선형종속이다.

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\2\\3\end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix}1\\0\\1\end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix}3\\2\\4\end{bmatrix} \quad \boldsymbol{v}_4 = \begin{bmatrix}0\\0\\1\end{bmatrix} \tag{A.8}$$

$\boldsymbol{v}_3$ 가 다음과 같이 다른 벡터의 선형결합으로 나타난는 것을 보여준다.

$$\boldsymbol{v}_3 = \begin{bmatrix}3\\2\\4\end{bmatrix} = (1)\boldsymbol{v}_1 + (2)\boldsymbol{v}_2 + (-1)\boldsymbol{v}_4 = (1)\begin{bmatrix}1\\2\\3\end{bmatrix} + (2)\begin{bmatrix}1\\0\\1\end{bmatrix} + (-1)\begin{bmatrix}0\\0\\1\end{bmatrix}$$

식 A.8 와 같이 3차원 벡터가 4개인 경우 벡터의 값에 관계없이 선형종속으로 나타난다. **n-차원 실수공간 R^n 에서 임의의 $n+1$ 개의 벡터는 항상 선형종속이다.**

A.13. 우드베리 공식

다음은 우드베리공식(Woodbury formula) 과 파생된 유용한 공식들이다.

$$(\boldsymbol{A}+\boldsymbol{UCV})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1}+\boldsymbol{VA}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VA}^{-1}$$

$$(\boldsymbol{I}+\boldsymbol{UCV})^{-1}=\boldsymbol{I}-\boldsymbol{U}(\boldsymbol{C}^{-1}+\boldsymbol{VU})^{-1}\boldsymbol{V}$$

$$(\boldsymbol{A}+\boldsymbol{uv}^t)^{-1}=\boldsymbol{A}^{-1}-\frac{\boldsymbol{A}^{-1}\boldsymbol{uv}^t\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{u}} \tag{A.9}$$

$$(a\boldsymbol{I}_n+b\mathbf{1}_n\mathbf{1}_n^t)^{-1}=\frac{1}{a}\left[\boldsymbol{I}_n-\frac{b}{a+nb}\mathbf{1}\mathbf{1}^t\right]$$

B. 고유값과 고유벡터

> Caution
>
> 고유값과 고유벡터는 정방행렬(square matrix)에 대해서만 정의된다.

B.1. 특성다항식

특성다항식(Characteristic polynomial)은 다음과 같이 정의된다

실수 $\lambda \in \mathbb{R}$ 와 정방행렬(square matrix) $A \in \mathbb{R}^{n\times n}$ 에 대하여

$$\begin{aligned} p_A(\lambda) &:= \mathsf{det}(A - \lambda I) \\ &= c_0 + c_1\lambda + c_2\lambda^2 + \cdots + c_{n-1}\lambda^{n-1} + (-1)^n\lambda^n, \end{aligned} \tag{B.1}$$

B.2. 고유값과 고유벡터

B.2.1. 정의

n-차원 정방행렬 $\boldsymbol{A}$ 이 있을 때, 다음 식을 만족하는 λ 와 벡터 $\boldsymbol{x}$가 존재하면 λ 를 행렬 $\boldsymbol{A}$ 의 고유값(eigenvalue), $\boldsymbol{x}$ 를 행렬 $\boldsymbol{A}$ 의 고유벡터(eigenvector)라고 한다 (부교재 definition 4.6)

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \tag{B.2}$$

- 고유벡터는 유일하지 않다. 즉, 벡터 $\boldsymbol{x}$ 가 고유벡터이면 $c\boldsymbol{x}$ 도 고유벡터이다.

$$\boldsymbol{A}(c\boldsymbol{x}) = c\boldsymbol{A}\boldsymbol{x} = c\lambda\boldsymbol{x} = \lambda(c\boldsymbol{x})$$

B.2.2. 계산

다음 4개의 문장은 동치이다

- λ 는 행렬 $\boldsymbol{A}$ 의 고유값이다.
- 방정식 $(\boldsymbol{A}-\lambda\boldsymbol{I})\boldsymbol{x}=\boldsymbol{0}$ 은 영벡터이외의 해 $\boldsymbol{x}$ 를 가진다(nontrivial solution)
- 행렬 $\boldsymbol{A}-\lambda\boldsymbol{I}$ 의 행렬식이 0 이다.

$$\mathsf{det}(\boldsymbol{A}-\lambda\boldsymbol{I})=0 \tag{B.3}$$

- 행렬 $\boldsymbol{A}-\lambda\boldsymbol{I}$ 의 rank가 n 보다 작다.

위에서 행렬식이 0 인 방정식 식 B.3 을 푸는 것은 특성방정식 식 B.1 이 0 인 방정식을 을 푸는 것과 동일하다.

Exercise B.1. 다음과 같은 2×2 행렬의 고유값과 고유행렬을 구해보자.

$$\boldsymbol{A}=\begin{bmatrix}4 & 2\\ 1 & 3\end{bmatrix}$$

다음과 같이 행렬 $\boldsymbol{A}$ 에 대한 특성다항식을 이용하여 고유값을 구할 수 있다.

$$\begin{aligned}p_{\boldsymbol{A}}(\lambda) &= \mathsf{det}(\boldsymbol{A}-\lambda\boldsymbol{I})\\ &= \mathsf{det}\left(\begin{bmatrix}4 & 2\\ 1 & 3\end{bmatrix}-\begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\right)=\begin{vmatrix}4-\lambda & 2\\ 1 & 3-\lambda\end{vmatrix}\\ &= (4-\lambda)(3-\lambda)-(2)(1)\\ &= (2-\lambda)(5-\lambda)\end{aligned}$$

위의 방정식에서 λ에 대한 다항식 $p_{\boldsymbol{A}}(\lambda)=0$ 의 근 을 구하면 고유값을 구할 수 있다. 따라서 행렬 $\boldsymbol{A}$ 의 고유값은 $\lambda_1=2$ 와 $\lambda_2=5$ 이다.

이제 각각의 고유값에 대한 고유행렬을 다음과 같이 고유벡터의 정의 식 B.2 에 의하여 구해보자.

$$\begin{bmatrix}4 & 2\\ 1 & 3\end{bmatrix}\boldsymbol{x}=\lambda\boldsymbol{x}\quad\rightarrow\quad\begin{bmatrix}4-\lambda & 2\\ 1 & 3-\lambda\end{bmatrix}\boldsymbol{x}=\boldsymbol{0}$$

먼저, 고유값 $\lambda_2=5$ 고유행렬은 다음과 같이 정의된다.

$$\begin{bmatrix}4-5 & 2\\ 1 & 3-5\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}-1 & 2\\ 1 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\boldsymbol{0}\quad\rightarrow x_1-2x_2=0$$

이제 위의 방정식을 만족하는 고유벡터는 다음과 같이 구할 수 있다.

$$\boldsymbol{x}_2=\begin{bmatrix}2\\ 1\end{bmatrix}$$

유의할 점은 고유벡터는 방정식을 만족하는 무수히 많은 벡터 중에 하나의 예일 뿐이다. 예를 들어 길이가 1 인 단위벡터(unit vector)인 고유벡터를 구하고 싶다면 위의 벡터를 길이가 1인 단위벡터로 바꾸면 된다.

$$\boldsymbol{x}_2 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$$

또한 $\lambda_1 = 2$ 에 대한 고유벡터는 다음과 같이 단위벡터로 구할 수 있다.

$$\begin{bmatrix} 4-2 & 2 \\ 1 & 3-2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \boldsymbol{x} = \boldsymbol{0} \quad \rightarrow x_1 + x_2 = 0$$

$$\boldsymbol{x}_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

■

B.2.3. 중복도와 고유공간

- 대수적 중복도(algebraic multiplicity) 는 특성다항식 식 B.1 이 0인 방정식을 푸는 경우 다항식에서 고유값이 중근(multiple root)의 해로 나타나는 차수를 의미한다.
- 기하적 중복도(geometric multiplicity) 는 고유값에 대응하는 고유벡터들 중 선형독립인 고유벡터들의 최대 개수를 의미한다.
- 고유 공간(eigenspace)은 고유값에 대응하는 고유벡터들이 생성하는 벡터공간을 의미한다.

Exercise B.2. 3차원 행렬 $\boldsymbol{A}$ 가 다음과 같을 때

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}$$

행렬 $\boldsymbol{A}$의 특성다항식은 다음과 같다.

$$\mathsf{det}(\lambda \boldsymbol{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda & 0 & 2 \\ -1 & \lambda - 2 & -1 \\ -1 & 0 & \lambda - 3 \end{vmatrix} = (\lambda - 1)(\lambda - 2)^2$$

참고로 특성방정식을 푸는 경우, 방정식 $\mathsf{det}(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$ 이나 $\mathsf{det}(\lambda \boldsymbol{I} - \boldsymbol{A}) = 0$ 중 어느 것을 사용해도 상관없다.

위의 식에서 3차원 행렬의 행렬식은 다음과 같이 구할 수 있다. 첫 번째 행을 기준으로 전개하면

$$\begin{aligned}\det(\lambda \boldsymbol{I} - \boldsymbol{A}) &= -\lambda \cdot \begin{vmatrix} 2-\lambda & 1 \\ 0 & 3-\lambda \end{vmatrix} + 0 \cdot \begin{vmatrix} 1 & 1 \\ 1 & 3-\lambda \end{vmatrix} + (-2) \cdot \begin{vmatrix} 1 & 2-\lambda \\ 1 & 0 \end{vmatrix} \\ &= -\lambda[(2-\lambda)(3-\lambda) - 0] - 2[(1)(0) - (1)(2-\lambda)] \\ &= (\lambda - 1)(\lambda - 2)^2\end{aligned}$$

첫번째 고유값은 $\lambda_1 = 1$ 이다. 고유벡터를 구하기 위해서는 다음과 같은 방정식을 풀면 된다.

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$$

위의 방정식을 풀면

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 1 & 0 & 2 \\ -1 & -1 & -1 \\ -1 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

아래와 같이 간단히 할 수 있으며

$$x_1 = -2x_3, \quad x_2 = x_3$$

다음과 같은 고유값과 고유벡터를 얻을 수 있다.

$$\lambda_1 = 1 \quad \rightarrow \quad \boldsymbol{x}_1 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

첫번째 고유값은 $\lambda_1 = 1$ 이며 대수적 중복도는 1이고 기하적 중복도도 1이다. 기하적 중복도도 1이란 의미는 고유벡터가 선형독립인 1개의 벡터로 이루어져 있다는 의미이다. 이 경우 고유공간 E_1 은 한 개의 고유벡터 $\boldsymbol{x}_1$ 이 생성하는 부분공간을 의미한다.

$$E_1 = \text{span}\left\{ \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} \right\}$$

다음으로 두번째 고유값에 대한 방정식 $(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$ 을 풀면 다음과 같다.

$$(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (2\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 2 & 0 & 2 \\ -1 & 0 & -1 \\ -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

이 방정식은 아래와 같이 간단히 할 수 있으며

$$x_1 = -x_3$$

주어진 방정식이 하나이기 때문에 다음과 같이 서로 선형독립인 두 개의 고유벡터를 얻을 수 있다.

$$\lambda_2 = 2 \quad \rightarrow \quad \boldsymbol{x}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \quad \boldsymbol{x}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

위에서 **두번째 고유값은 $\lambda_2 = 2$ 이며 대수적 중복도는 2이다.** 또한 **선형독립인 2개의 고유벡터를 구할 수 있으므로 기하적 중복도는 2이다.**

이 경우 E_2 는 두 개의 고유벡터 $\boldsymbol{x}_2, \boldsymbol{x}_3$ 가 생성하는 부분공간을 의미한다.

$$E_2 = \text{span} \left\{ \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

■

이제 대수적 중복도와 기하적 중복도가 다른 경우에 대한 예제를 들어보자.

Exercise B.3. 3차원 행렬 $\boldsymbol{A}$ 가 다음과 같을 때

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 2 \\ -1 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix}$$

행렬 $\boldsymbol{A}$의 특성다항식은 다음과 같다.

$$\det(\lambda \boldsymbol{I} - \boldsymbol{A}) = \begin{vmatrix} \lambda - 1 & 0 & -2 \\ 1 & \lambda - 1 & -3 \\ 0 & 0 & \lambda - 2 \end{vmatrix} = (\lambda - 1)^2 (\lambda - 2)$$

첫번째 고유값은 $\lambda_1 = 1$ 이다. 고유벡터를 구하기 위해서는 다음과 같은 방정식을 풀면 된다.

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$$

위의 방정식을 풀면

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 0 & -3 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

아래와 같이 간단히 할 수 있으며

$$x_1 = x_3 = 0$$

다음과 같은 하나의 고유벡터를 얻을 수 있다.

$$\lambda_1 = 1 \quad \rightarrow \quad x_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

첫번째 고유값은 $\lambda_1 = 1$ 이며 대수적 중복도는 2이지만 기하적 중복도는 1이다. 이 경우 고유공간 E_1 은 한 개의 고유벡터 $\boldsymbol{x}_1$ 이 생성하는 부분공간을 의미한다.

$$E_1 = \text{span} \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

다음으로 두번째 고유값에 대한 방정식 $(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$ 을 풀면 다음과 같다.

$$(\lambda_2 \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = (2\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & -3 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

이 방정식은 아래와 같이 간단히 할 수 있으며

$$x_1 = -2x_3, \quad x_2 = 5x_3$$

다음과 같은 한 개의 고유벡터를 얻을 수 있다.

$$\lambda_2 = 2 \quad \rightarrow \quad x_2 = \begin{bmatrix} -2 \\ 5 \\ 1 \end{bmatrix}$$

위에서 **두번째 고유값은 $\lambda_2 = 2$ 이며 대수적 중복도는 1이다.** 또한 **선형독립인 1개의 고유벡터를 구할 수 있으므로 기하적 중복도는 1이다.**

이 경우 E_2 는 한 개의 고유벡터 $\boldsymbol{x}_2$ 가 생성하는 부분공간을 의미한다.

$$E_2 = \text{span}\left\{ \begin{bmatrix} -2 \\ 5 \\ 1 \end{bmatrix} \right\}$$

■

B.3. 대칭행렬의 대각화

Exercise B.4. 이제 대칭행렬에 대한 고유값과 고유행렬을 구해보자. 대칭행렬은 고유값이 실수이고 서로 다르며, 서로 직교하는 고유벡터를 가진다.

$$\boldsymbol{A} = \tfrac{1}{2} \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$$

행렬 $\boldsymbol{A}$ 의 특성다항식은 다음과 같이 구할 수 있다.

$$\begin{aligned} \det(\boldsymbol{A} - \lambda \boldsymbol{I}) &= \det\left(\begin{bmatrix} \frac{5}{2} - \lambda & -1 \\ -1 & \frac{5}{2} - \lambda \end{bmatrix} \right) \\ &= \left(\frac{5}{2} - \lambda \right)^2 - 1 = \lambda^2 - 5\lambda + \frac{21}{4} = \left(\lambda - \frac{7}{2} \right) \left(\lambda - \frac{3}{2} \right) \end{aligned}$$

따라서 행렬 $\boldsymbol{A}$ 의 고유값은 각각 $\lambda_1 = \frac{7}{2}$ 과 $\lambda_2 = \frac{3}{2}$ 이며 대응하는 고유벡터 $\boldsymbol{p}_1$ 과 $\boldsymbol{p}_2$ 는 다음과 같이 구할 수 있다.

$$\boldsymbol{A}\boldsymbol{p}_1 = \frac{7}{2}\boldsymbol{p}_1, \quad \boldsymbol{A}\boldsymbol{p}_2 = \frac{3}{2}\boldsymbol{p}_2$$

위의 고유빅터에 대한 방정식을 풀어서 길이가 1인 고유벡터를 구하면 다음과 같다.

$$\boldsymbol{p}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{p}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

이제 고유벡터는 서로 직교하는 단위벡터임을 알 수 있다.

$$\boldsymbol{p}_1^t \boldsymbol{p}_2 = 0$$

이제 고유값과 고유벡터를 구했으니 대칭행렬 $\boldsymbol{A}$ 를 대각화 해보자. 대칭행렬 $\boldsymbol{A}$ 는 다음과 같이 대각화 할 수 있다. 먼저 두 고유벡터를 열벡터로 하는 행렬 $\boldsymbol{P}$ 를 정의하자. 행렬 $\boldsymbol{P}$ 는 서로 직교하는 벡터로 구성되었으므로 직교행렬이다.

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \rightarrow \quad \boldsymbol{P}\boldsymbol{P}^t = \boldsymbol{P}^t\boldsymbol{P} = \boldsymbol{I}$$

이제 $\boldsymbol{P}$ 의 역행렬은 $\boldsymbol{P}^t$ 이므로 다음과 같이 대칭행렬 $\boldsymbol{A}$의 대각화를 유도할 수 있다. 이러한 대각화를 행렬의 스펙트럼분해(Spectral Decomposition)이라고 부른다.

$$\boldsymbol{P}^{-t}\boldsymbol{A}\boldsymbol{P} = \begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} = \boldsymbol{D}. \tag{B.4}$$

위의 식은 다음과 같이 쓸수 있다.

$$\underbrace{\frac{1}{2}\begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}}_{\boldsymbol{A}} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{\boldsymbol{P}} \underbrace{\begin{bmatrix} \frac{7}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_{\boldsymbol{P}^t}$$

또한 위의 식은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \boldsymbol{A} &= \frac{1}{2}\begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix} \\ &= \lambda_1 \boldsymbol{p}_1 \boldsymbol{p}_1^t + \lambda_2 \boldsymbol{p}_2 \boldsymbol{p}_2^t \\ &= \tfrac{7}{4}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix} + \frac{3}{4}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \end{bmatrix} \\ &= \tfrac{7}{4}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{3}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{aligned}$$

■

C. 행렬의 분해

C.1. Gram–Schmidt 방법

서로 독립인 n차원의 벡터들이 p개 있을때

$$\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$$

이들이 만드는 열공간을 C 라고 하자.

$$\begin{aligned} C &= span\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p\} \\ &= \{\ c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_p\boldsymbol{a}_p \mid \text{ all possible real values of } c_1, c_2, \dots, c_p\ \} \end{aligned}$$

이제 우리는 위와 동일한 열공간 C 만드는 정규직교 벡터들을 찾는 방법을 알아보고자 한다.

$$\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p \quad \text{where } \boldsymbol{q}_i^t\boldsymbol{q}_j = 0,\ \ \boldsymbol{q}_i^t\boldsymbol{q}_i = 1$$

그리고

$$C = span\{\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p\} = span\{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p\} \tag{C.1}$$

이제 앞 절의 벡터의 사영에 대한 결과를 사용하여 다음과 같은 직교하는 p 개의 벡터들을 축차적으로 만들어 보자.

$$\begin{aligned} \tilde{\boldsymbol{q}}_1 &= \boldsymbol{a}_1 \\ \tilde{\boldsymbol{q}}_2 &= \boldsymbol{a}_2 - proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_2) \\ \tilde{\boldsymbol{q}}_3 &= \boldsymbol{a}_3 - proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_3) - proj_{\tilde{\boldsymbol{q}}_2}(\boldsymbol{a}_3) \\ \tilde{\boldsymbol{q}}_4 &= \boldsymbol{a}_4 - proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_4) - proj_{\tilde{\boldsymbol{q}}_2}(\boldsymbol{a}_4) - proj_{\tilde{\boldsymbol{q}}_3}(\boldsymbol{a}_4) \\ &\dots \\ \tilde{\boldsymbol{q}}_p &= \boldsymbol{a}_p - \sum_{k=1}^{p} proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_p) \end{aligned}$$

축차적으로 만든 벡터들을 정규벡터로 만들면 원래의 벡터들 $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$이 생성하는 동일한 열공간을 만드는 정규직교 벡터 $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p$를 만들 수 있다.

$$\boldsymbol{q}_i = \tilde{\boldsymbol{q}}_i / \left\|\tilde{\boldsymbol{q}}_i\right\|, \quad i = 1, 2, \dots, p \tag{C.2}$$

Gram–Schmidt 방법으로 만든 벡터들이 직교하는 것은 다음과 같이 증명할 수 있다. 먼저 **?@eq-proofortho** 에 의하여 $\tilde{\boldsymbol{q}}_1$ 과 $\tilde{\boldsymbol{q}}_2$는 직교한다. 이제 임의의 i에 대하여 $\tilde{\boldsymbol{q}}_1, \tilde{\boldsymbol{q}}_2, \cdots, \tilde{\boldsymbol{q}}_{i-1}$ 벡터들이 직교한다고 가정하자. 모든 $1 \le j \le i-1$ 에 대하여

$$\begin{aligned}
\tilde{\boldsymbol{q}}_j^t \tilde{\boldsymbol{q}}_i &= \tilde{\boldsymbol{q}}_j^t \left[\boldsymbol{a}_i - \sum_{k=1}^{i-1} proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i) \right] \\
&= \tilde{\boldsymbol{q}}_j^t \left[\boldsymbol{a}_i - proj_{\tilde{\boldsymbol{q}}_j}(\boldsymbol{a}_i) \right] - \left[\sum_{\substack{1 \le k \le i-1 \\ k \ne j}} \tilde{\boldsymbol{q}}_j^t \, proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i) \right] \\
&= 0 + 0
\end{aligned}$$

위에서 마지막 단계의 직교성은 다음과 같은 사실로 부터 유도된다.

- $\boldsymbol{a}_i - proj_{\tilde{\boldsymbol{q}}_j}(\boldsymbol{a}_i)$는 $\tilde{\boldsymbol{q}}_j^t$와 직교한다.
- 가정에 의하여 $\tilde{\boldsymbol{q}}_1, \tilde{\boldsymbol{q}}_2, \cdots, \tilde{\boldsymbol{q}}_{i-1}$ 는 직교하고 $proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i)$ 는 $\tilde{\boldsymbol{q}}_k$ 와 같은 방향을 가진다.

$$\tilde{\boldsymbol{q}}_j^t proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_i) = 0 \quad \text{for } 1 \le j, k \le i-1, k \ne j$$

식 C.2 과 식 C.2 의 알고리즘을 Gram–Schmidt 방법이라고 부른다. 위의 두 식에 의한 알고리즘을 다음과 같은 사실을 이용하면 좀 더 간단한 방법의 알고리즘이 나온다.

$$proj_{\tilde{\boldsymbol{q}}_k}(\boldsymbol{a}_l) = \frac{\boldsymbol{a}_l^t \tilde{\boldsymbol{q}}_k}{\tilde{\boldsymbol{q}}_k^t \tilde{\boldsymbol{q}}_k} \tilde{\boldsymbol{q}}_k = \frac{\boldsymbol{a}_l^t \tilde{\boldsymbol{q}}_k}{\left\|\tilde{\boldsymbol{q}}_k\right\|^2} \tilde{\boldsymbol{q}}_k = (\boldsymbol{a}_l^t \boldsymbol{q}_k) \boldsymbol{q}_k$$

1. p개의 벡터 $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$에 대하여
2. for $i = 1, 2, \dots, p$
    - $\tilde{\boldsymbol{q}}_i = \boldsymbol{a}_i - (\boldsymbol{q}_1^t \boldsymbol{a}_i)\boldsymbol{q}_1 - \cdots - (\boldsymbol{q}_{i-1}^t \boldsymbol{a}_i)\boldsymbol{q}_{i-1}$ (직교화)
    - $\boldsymbol{q}_i = \tilde{\boldsymbol{q}}_i / \left\|\boldsymbol{q}_i\right\|$ (정규화)

다음은 Gram–Schmidt 방법을 설명한 그림이다.

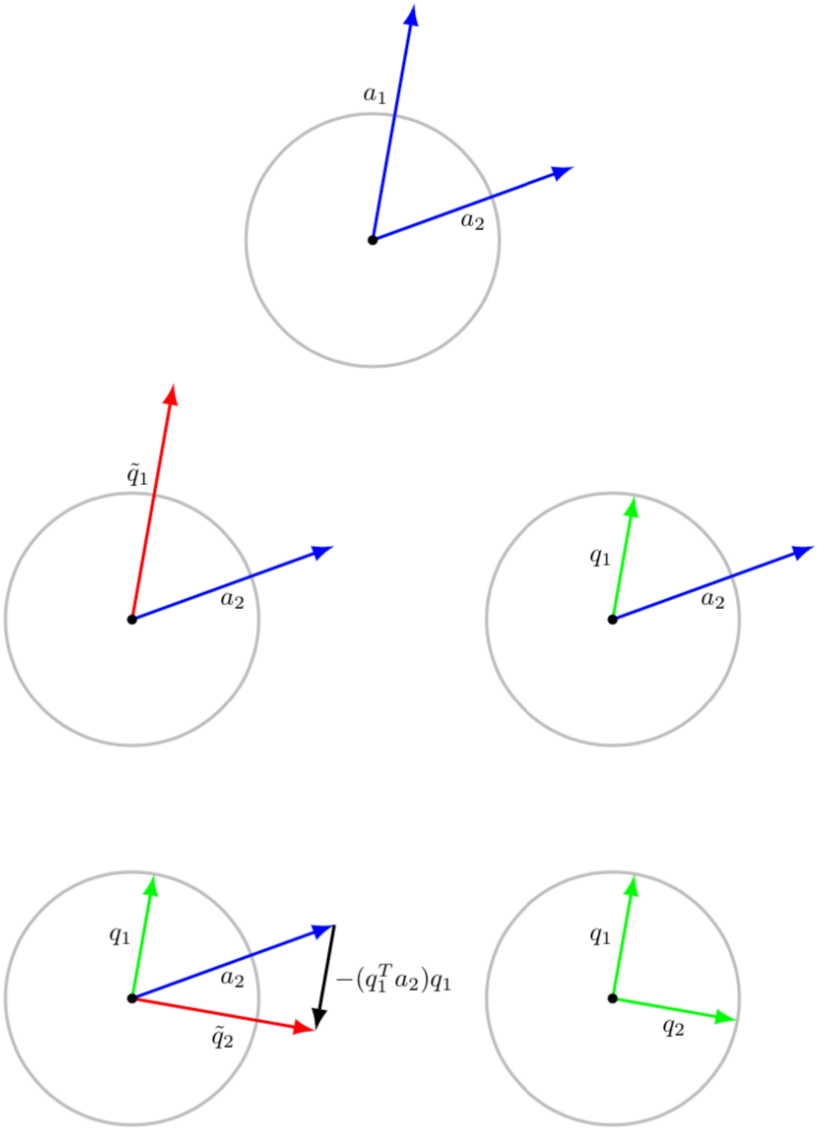


Figure 5.3 Gram–Schmidt algorithm applied to two 2-vectors a_1, a_2. *Top.* The original vectors a_1 and a_2. The gray circle shows the points with norm one. *Middle left.* The orthogonalization step in the first iteration yields $\tilde{q}_1 = a_1$. *Middle right.* The normalization step in the first iteration scales $\tilde{q}_1$ to have norm one, which yields q_1. *Bottom left.* The orthogonalization step in the second iteration subtracts a multiple of q_1 to yield the vector $\tilde{q}_2$, which is orthogonal to q_1. *Bottom right.* The normalization step in the second iteration scales $\tilde{q}_2$ to have norm one, which yields q_2.

그림 C.1.: Gram–Schmidt 방법(출처:Introduction to Applied Linear Algebra by Boyd and Vandenberghe, 2019)

C.2. LU 분해

정방행렬 $\boldsymbol{A}$를 다음과 같이 하삼각행렬 $\boldsymbol{L}$과 상삼각행렬 $\boldsymbol{U}$의 곱으로 나타내는 것을 LU 분해라고 한다.

$$\boldsymbol{A} = \boldsymbol{LU}$$

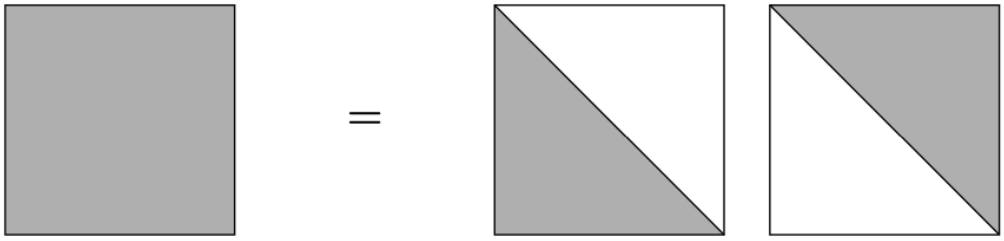


그림 C.2.: LU 분해

이러한 LU 분해는 행렬 $\boldsymbol{A}$에 행연산을 적용하여 쉽게 구할 수 있다. 예를 들어 위에서 고려한 2×2 행렬에 행연산을 적용하여 대각원소 아래를 0으로 만들면 LU 분해를 쉽게 유도할 수 있다.

$$\begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}$$

따라서

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix} = \boldsymbol{LU}$$

C.3. QR 분해

식 C.2 과 식 C.2 에 주어진 Gram–Schmidt 방법을 원래 벡터들 $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p$에 대하여 다시 다음과 같이 나타낼 수 있다.

$$\begin{aligned}
\boldsymbol{a}_1 &= \tilde{\boldsymbol{q}}_1 \\
&= \|\tilde{\boldsymbol{q}}_1\| \boldsymbol{q}_1 \\
\boldsymbol{a}_2 &= \tilde{\boldsymbol{q}}_2 + proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_2) \\
&= \tilde{\boldsymbol{q}}_2 + \frac{\boldsymbol{a}_2^t \tilde{\boldsymbol{q}}_1}{\tilde{\boldsymbol{q}}_1^t \tilde{\boldsymbol{q}}_1} \tilde{\boldsymbol{q}}_1 \\
&= (\boldsymbol{a}_2^t \boldsymbol{q}_1)\boldsymbol{q}_1 + \|\tilde{\boldsymbol{q}}_2\| \boldsymbol{q}_2 \\
\boldsymbol{a}_3 &= \tilde{\boldsymbol{q}}_3 + proj_{\tilde{\boldsymbol{q}}_1}(\boldsymbol{a}_3) + proj_{\tilde{\boldsymbol{q}}_2}(\boldsymbol{a}_3) \\
&= \tilde{\boldsymbol{q}}_3 + \frac{\boldsymbol{a}_3^t \tilde{\boldsymbol{q}}_1}{\tilde{\boldsymbol{q}}_1^t \tilde{\boldsymbol{q}}_1} \tilde{\boldsymbol{q}}_1 + \frac{\boldsymbol{a}_3^t \tilde{\boldsymbol{q}}_2}{\tilde{\boldsymbol{q}}_2^t \tilde{\boldsymbol{q}}_2} \tilde{\boldsymbol{q}}_2 \\
&= (\boldsymbol{a}_3^t \boldsymbol{q}_1)\boldsymbol{q}_1 + (\boldsymbol{a}_3^t \boldsymbol{q}_2)\boldsymbol{q}_2 + \|\tilde{\boldsymbol{q}}_3\| \boldsymbol{q}_3 \\
&\dots \\
\boldsymbol{a}_p &= (\boldsymbol{a}_p^t \boldsymbol{q}_1)\boldsymbol{q}_1 + (\boldsymbol{a}_p^t \boldsymbol{q}_2)\boldsymbol{q}_2 + \cdots + (\boldsymbol{a}_p^t \boldsymbol{q}_{p-1})\boldsymbol{q}_{p-1} + \|\tilde{\boldsymbol{q}}_p\| \boldsymbol{q}_p
\end{aligned}$$

즉 위의 축차식을 보면 원래 벡터 $\boldsymbol{a}_i$ 는 Gram–Schmidt 방법으로 구한 정규직교벡터 $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p$ 의 선형 조합으로 나타낼 수 있다.

이제 Gram–Schmidt 방법으로 구한 정규직교벡터들 $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_p$ 을 모아놓은 행렬을 $\boldsymbol{Q}$ 라고 하고 위에서 $\boldsymbol{a}_i$ 들이 직교행렬의 선형조합으로 표시될때 계수들을 모아놓는 상삼각행렬을 $\boldsymbol{R}$ 이라고 하자. 그려면 다음과 같은 QR 분해가 주어진다.

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} \tag{C.3}$$

여기서

$$\boldsymbol{Q} = [\boldsymbol{q}_1 \;\; \boldsymbol{q}_2 \;\; \cdots \;\; \boldsymbol{q}_p], \quad \boldsymbol{Q}^t\boldsymbol{Q} = \boldsymbol{I}$$

$$\boldsymbol{R} = \begin{bmatrix} \|\tilde{\boldsymbol{q}}_1\| & \boldsymbol{a}_2^t\boldsymbol{q}_1 & \boldsymbol{a}_3^t\boldsymbol{q}_1 & \cdots & \boldsymbol{a}_p^t\boldsymbol{q}_1 \\ 0 & \|\tilde{\boldsymbol{q}}_2\| & \boldsymbol{a}_3^t\boldsymbol{q}_2 & \cdots & \boldsymbol{a}_p^t\boldsymbol{q}_2 \\ 0 & 0 & \|\tilde{\boldsymbol{q}}_3\| & \cdots & \boldsymbol{a}_p^t\boldsymbol{q}_3 \\ & & & \cdots & \\ 0 & 0 & 0 & \cdots & \|\tilde{\boldsymbol{q}}_p\| \end{bmatrix}$$

이제 Gram–Schmidt 방법과 QR 분해를 실제 예제를 통하여 구해보자

아래와 같이 4차원 벡터 3개가 있다.

$$\boldsymbol{a}_1 = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \quad \boldsymbol{a}_2 = \begin{bmatrix} -1 \\ 3 \\ -1 \\ 3 \end{bmatrix} \quad \boldsymbol{a}_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \end{bmatrix} \tag{C.4}$$

위의 벡터 $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$에 대하여 Gram–Schmidt 방법을 적용해보자.

1. $i = 1$. 먼저 $\|\tilde{\boldsymbol{q}}_1\| = \|\boldsymbol{a}_1\| = 2$이므로 첫번째 벡터 $\boldsymbol{q}_1$를 만든다.

$$\boldsymbol{q}_1 = \tilde{\boldsymbol{q}}_1 / \|\tilde{\boldsymbol{q}}_1\| = \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix}$$

2. $i = 2$. 이제 두번째 직교벡터 $\boldsymbol{q}_2$를 만들자. $\boldsymbol{q}_1^t\boldsymbol{a}_2 = 4$이므로

$$\tilde{\boldsymbol{q}_2} = \boldsymbol{a}_2 - (\boldsymbol{q}_1^t \boldsymbol{a}_2)\boldsymbol{q}_1 = \begin{bmatrix} -1 \\ 3 \\ -1 \\ 3 \end{bmatrix} - 4 \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

그리고 $\|\tilde{\boldsymbol{q}}_2\| = 2$이므로

$$\boldsymbol{q}_2 = \tilde{\boldsymbol{q}}_2 / \|\tilde{\boldsymbol{q}}_2\| = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

3. $i = 3$ 마지막으로 $\boldsymbol{q}_1^t \boldsymbol{a}_3 = 2$, $\boldsymbol{q}_2^t \boldsymbol{a}_3 = 8$ 이므로

$$\tilde{\boldsymbol{q}_3} = \boldsymbol{a}_3 - (\boldsymbol{q}_1^t \boldsymbol{a}_3)\boldsymbol{q}_1 - (\boldsymbol{q}_2^t \boldsymbol{a}_3)\boldsymbol{q}_2 = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \end{bmatrix} - 2 \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix} - 8 \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix}$$

또한 $\|\tilde{\boldsymbol{q}}_3\| = 4$이므로

$$\boldsymbol{q}_3 = \tilde{\boldsymbol{q}}_3 / \|\tilde{\boldsymbol{q}}_3\| = \begin{bmatrix} -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

따라서 Gram–Schmidt 방법으로 만든 정규직교벡터는 다음과 같다.

$$\boldsymbol{q}_1 = \begin{bmatrix} -1/2 \\ 1/2 \\ -1/2 \\ 1/2 \end{bmatrix} \quad \boldsymbol{q}_2 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} \quad \boldsymbol{q}_3 = \begin{bmatrix} -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

이제 위에서 구한 Gram-Schmidt 방법으로 얻은 결과를 이용하여 QR 분해를 구해보자.

식 C.4 에서 주어진 백터들을 열로 가진 행렬 $\boldsymbol{A}$의 QR분해를 구해보자.

$$\boldsymbol{A} = \begin{bmatrix} -1 & -1 & 1 \\ 1 & 3 & 3 \\ -1 & -1 & 5 \\ 1 & 3 & 7 \end{bmatrix}$$

앞의 예제에서 구한 직교벡터를 그대로 이용하면 $\boldsymbol{Q}$는 쉽게 구해진다.

$$\boldsymbol{Q} = \begin{bmatrix} -1/2 & 1/2 & -1/2 \\ 1/2 & 1/2 & -1/2 \\ -1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \end{bmatrix}$$

또한 식 C.3 에 주어진 공식을 이용하면 행렬 $\boldsymbol{R}$은 다음과 같이 구할 수 있다.

$$\boldsymbol{R} = \begin{bmatrix} \|\tilde{\boldsymbol{q}}_1\| & \boldsymbol{a}_2^t \boldsymbol{q}_1 & \boldsymbol{a}_3^t \boldsymbol{q}_1 \\ 0 & \|\tilde{\boldsymbol{q}}_2\| & \boldsymbol{a}_3^t \boldsymbol{q}_2 \\ 0 & 0 & \|\tilde{\boldsymbol{q}}_3\| \end{bmatrix} = \begin{bmatrix} 2 & 4 & 2 \\ 0 & 2 & 8 \\ 0 & 0 & 4 \end{bmatrix}$$

C.4. SVD 분해

C.4.1. 특이값과 특이벡터

고유값과 고유벡터는 정방행렬인 경우 정의되는 것으로서 행렬이 정방행렬이 아닌 경우에는 구할 수 없다. 이제 고유값과 유사한 성질을 가지는 특이값을 일반행렬에서 정의해보자.

$\boldsymbol{A}$가 $m \times n$ 일반행렬이라고 가정하고 그 계수 r이라고 하자 ($r(\boldsymbol{A}) = r$). 이제 서로 직교하는 n-차원의 벡터들의 집합 $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n$과 다른 직교하는 m-차원의 벡터들의 집합 $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m$을 생각하자.

행렬 $\boldsymbol{A}$의 특이값(singular values) $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$과 왼쪽 특이벡터(left singular vectors) $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m$ 그리고 오른쪽 특이벡터(right singular vectors) $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n$ 는 다음과 같은 성질을 만족한다.

$$\boldsymbol{A}\boldsymbol{v}_1 = \sigma_1 \boldsymbol{u}_1, \quad \boldsymbol{A}\boldsymbol{v}_2 = \sigma_2 \boldsymbol{u}_2, \quad \dots \quad \boldsymbol{A}\boldsymbol{v}_r = \sigma_r \boldsymbol{u}_r, \quad \boldsymbol{A}\boldsymbol{v}_{r+1} = \boldsymbol{0}, \quad \dots, \quad \boldsymbol{A}\boldsymbol{v}_n = \boldsymbol{0} \tag{C.5}$$

$n \times n$ 정방행렬 $\boldsymbol{V}$와 $m \times m$ 정방행렬 $\boldsymbol{U}$ 를 각각 서로 직교하는 정규벡터 $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n$ 과 $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m$ 으로 구성되는 직교행렬이라고 하자.

$$\boldsymbol{V} = [\boldsymbol{v}_1 \; \boldsymbol{v}_2 \; \dots \; \boldsymbol{v}_n], \quad \boldsymbol{U} = [\boldsymbol{u}_1 \; \boldsymbol{u}_2 \; \dots \; \boldsymbol{u}_m]$$

식 C.5 에 나타난 관계를 행렬 $\boldsymbol{V}$와 $\boldsymbol{U}$로 나타내면 다음과 같이 표현할 수 있다.

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \tag{C.6}$$

위에서 $m \times n$ 행렬 $\boldsymbol{\Sigma}$는 다음과 같은 형태를 가진다.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\Sigma}_r = \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & & \ddots \\ 0 & & \sigma_r \end{bmatrix}$$

C.4.2. SVD 분해

이제 행렬 $\boldsymbol{V}$가 직교행렬을 이용하면 다음과 같은 SVD 분해(singular value decomposition; 특이값 분해)을 정의할 수 있다.

$$\underset{m\times n}{\boldsymbol{A}} = \underset{m\times m}{\boldsymbol{U}} \; \underset{m\times n}{\boldsymbol{\Sigma}} \; \underset{n\times n}{\boldsymbol{V}^t} \tag{C.7}$$

위의 식 C.7 을 전개하면 다음과 같이 계수가 1인 행렬 $\boldsymbol{u}_k\boldsymbol{v}_k^t$ 들의 선형조합으로 행렬 $\boldsymbol{A}$를 나타낼 수 있다.

$$\boldsymbol{A} = \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^t + \sigma_2\boldsymbol{u}_2\boldsymbol{v}_2^t + \dots \sigma_r\boldsymbol{u}_r\boldsymbol{v}_r^t \tag{C.8}$$

또한 식 C.6 에서 $\boldsymbol{\Sigma}$에서 0이 되는 값을 제외하면 처음 r개의 요소들만 이루어진 부분으로만 축소된 SVD 분해를 구할 수 있다.

$$\boldsymbol{A}\boldsymbol{V}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}_r, \quad \boldsymbol{A}[\boldsymbol{v}_1\ \boldsymbol{v}_2\ \dots\ \boldsymbol{v}_r] = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \dots\ \boldsymbol{u}_r] \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & & \ddots \\ 0 & & \sigma_r \end{bmatrix} \tag{C.9}$$

위의 식 C.9 에서 주의할 점은 행렬 $\boldsymbol{V}_r$과 $\boldsymbol{U}_r$은 정방행렬이 아니고 직교행렬도 아니다. $\boldsymbol{V}_r^t\boldsymbol{V}_r = \boldsymbol{I}$ 와 $\boldsymbol{U}_r^t\boldsymbol{U}_r = \boldsymbol{I}$ 이 성립하지만 일반적으로 $\boldsymbol{V}_r\boldsymbol{V}_r^t \neq \boldsymbol{I}, \boldsymbol{U}_r\boldsymbol{U}_r^t \neq \boldsymbol{I}$이다.

C.4.3. 특이값과 특이벡터의 계산

$m \times n$ 행렬 $\boldsymbol{A}$의 SVD 분해 식 C.7 로 부터 행렬 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$를 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \boldsymbol{A}^t\boldsymbol{A} &= (\boldsymbol{V}\boldsymbol{\Sigma}^t\boldsymbol{U}^t)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t) = \boldsymbol{V}\boldsymbol{\Sigma}^t\boldsymbol{\Sigma}\boldsymbol{V}^t \\ \boldsymbol{A}\boldsymbol{A}^t &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t)(\boldsymbol{V}\boldsymbol{\Sigma}^t\boldsymbol{U}^t) = \boldsymbol{U}\boldsymbol{\Sigma}\Sigma^t\boldsymbol{U}^t \end{aligned} \tag{C.10}$$

위에서 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$는 모두 대칭행렬이지만 서로 차원이 다르다. 또한 식 C.10 을 보면 두 행렬이 모두 $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^t$의 형식으로 분해되는 것을 알 수 있다. 즉 다음과 같은 사실을 알 수 있다.

- $n \times n$ 비음정치행렬 $\boldsymbol{A}^t\boldsymbol{A}$의 고유벡터 행렬은 $\boldsymbol{V}$이다.
- $m \times m$ 비음정치행렬 $\boldsymbol{A}\boldsymbol{A}^t$의 고유벡터 행렬은 $\boldsymbol{U}$이다.
- 행렬 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$의 0이 아닌 고유값은 $\sigma_1^2, \sigma_2^2, \dots, \sigma_r^2$ 이다.

따라서 다음과 같은 방법으로 특이값과 특이벡터를 계산할 수 있다. 위의 방법은 두 행렬 $\boldsymbol{A}^t\boldsymbol{A}$와 $\boldsymbol{A}\boldsymbol{A}^t$를 모두 구하지 않고 $\boldsymbol{A}^t\boldsymbol{A}$의 고유값과 고유벡터만으로 SVD 분해를 구하는 방법이다 (만약 행렬 $\boldsymbol{A}$가 100000×5이라면 $\boldsymbol{A}\boldsymbol{A}^t$는 100000×100000이다!)

먼저 $\boldsymbol{A}^t\boldsymbol{A}$의 고유벡터 $\boldsymbol{v}_1, \dots, \boldsymbol{v}_r$을 다음과 같은 고유값과 고유벡터의 정의로 먼저 구한다.

$$\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k = \sigma_k^2 \boldsymbol{v}_k, \quad k = 1, 2, \dots, r \tag{C.11}$$

다음으로 다음의 식으로 $\boldsymbol{u}_1, \dots, \boldsymbol{u}_r$ 를 구한다.

$$\boldsymbol{u}_k = \frac{\boldsymbol{A}\boldsymbol{v}_k}{\sigma_k}, \quad k = 1, 2, \dots, r \tag{C.12}$$

식 C.12 에서 다음과 같이 $\boldsymbol{u}_k$가 행렬 $\boldsymbol{A}\boldsymbol{A}^t$ 의 고유벡터임을 확인할 수 있다.

$$\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{u}_k = \boldsymbol{A}\boldsymbol{A}^t\left(\frac{\boldsymbol{A}\boldsymbol{v}_k}{\sigma_k}\right) = \boldsymbol{A}\left(\frac{\sigma_k^2\boldsymbol{v}_k}{\sigma_k}\right) = \sigma_k^2\boldsymbol{u}_k$$

또한 식 C.11 에서 $\boldsymbol{v}_k$는 정규직교벡터이므로 다음과 같이 $\boldsymbol{u}_k$도 정규직교행려임을 보일 수 있다.

$$\boldsymbol{u}_k^t\boldsymbol{u}_l = \left(\frac{\boldsymbol{A}\boldsymbol{v}_k}{\sigma_k}\right)^t\left(\frac{\boldsymbol{A}\boldsymbol{v}_l}{\sigma_l}\right) = \frac{\boldsymbol{v}_k^t(\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{v}_l)}{\sigma_k\sigma_l} = \frac{\sigma_l}{\sigma_k}\boldsymbol{v}_k^t\boldsymbol{v}_l = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{if } k \neq l \end{cases}$$

위에서 구한 r개의 $\boldsymbol{v}_k$와 $\boldsymbol{u}_k$ 외에 $n-r$과 $m-r$ 개의 서로 직교하는 나머지 $\boldsymbol{v}$와 $\boldsymbol{u}$도 구할 수 있다.

C.4.4. SVD 분해의 기하학적 의미

다음은 SVD 분해의 기하학적 의미를 설명한 그림이다.

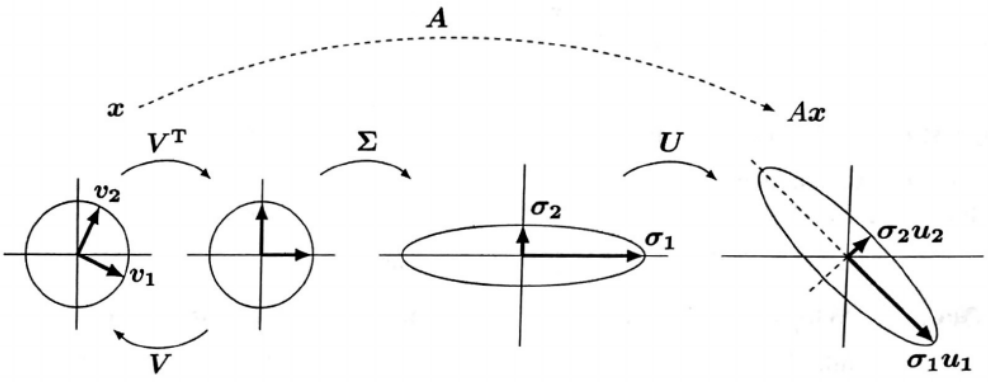


그림 C.3.: SVD 분해의 기하학적 의미

C.5. 양정치행렬

C.5.1. 이차형식

n-차원 벡터 $\boldsymbol{x}^t = [x_1, x_2, \dots, x_n]$과 대칭행렬 $\boldsymbol{A}$에 대하여 이차형식(quadratic form)은 다음과 같이 정의된다.

$$Q_A(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{C.13}$$

이차형식의 정의에서 반드시 행렬 $\boldsymbol{A}$를 대칭행렬로 정의하지 않아도 되지만 임의의 행렬에 대하여 이차형식의 값이 동일한 대칭행렬이 존재하기 때문에 정의에서 이차형식으로 국한하는 것이 일반적이다.

Definition C.1 (양정치 행렬)**.** 이차형식 $Q_A(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x}$가 영벡터가 아닌 모든 벡터 $\boldsymbol{x}$에 대하여 0 보다 크면, 즉

$$\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} > 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^n$$

$\boldsymbol{A}$를 양정치(positive definite)라고 부른다.

만약 이차형식 $Q_A(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x}$가 영벡터가 아닌 모든 벡터 $\boldsymbol{x}$에 대하여 0 보다 크거나 같다면

$$\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} \geq 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^n$$

$\boldsymbol{A}$를 양반정치(positive semi-definite)라고 부른다.

■

정칙행렬 $\boldsymbol{B}$에 대하여 다음과 같은 선형변환을 고려하자.

$$\boldsymbol{x} = \boldsymbol{B} \boldsymbol{y} \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{B}^{-1} \boldsymbol{x}$$

벡터 $\boldsymbol{x}$로 정의된 이차형식은 벡터 $\boldsymbol{y}$의 형태로 다음과 같이 변환할 수 있다.

$$Q(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = \boldsymbol{y}^t \boldsymbol{B}^t \boldsymbol{A} \boldsymbol{B} \boldsymbol{y} = Q^*(\boldsymbol{y})$$

이차형식의 성질은 정칙 선형변환에서 유지된다. 즉 행렬 $\boldsymbol{A}$가 양(반)정치 행렬이고 행렬 $\boldsymbol{B}$가 정칙행렬이면 행렬 $\boldsymbol{B}^t \boldsymbol{A} \boldsymbol{B}$도 양(반)정치 행렬이다.

C.5.2. 양정치행렬의 성질

양정치행렬 $\boldsymbol{A}$ 은 모든 고유값은 양수이다. 따라서 고유값과 고유벡터의 성질에 위해 유도된 식 B.4 에 의하여 다음과 같은 분해가 가능하다.

$$\begin{aligned}\boldsymbol{A} &= \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^t \\ &= \boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{P}^t \\ &= \left[\boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}^t\right]\left[\boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}^t\right] \\ &= \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\end{aligned} \tag{C.14}$$

위의 식에서 $\boldsymbol{D}^{1/2}$는 $\boldsymbol{D}$의 대각원소의 제곱근을 원소로 하는 대각행렬이다. 이는 양정치 행렬의 고유값이 모두 양수이기 때문에 가능하다.

예를 들어 다변량 확률 벡터의 공분산 행렬 $\boldsymbol{\Sigma}$ 또는 상관행렬 $\boldsymbol{R}$ 은 모두 양정치 행렬이다. 따라서 다음과 같은 분해가 가능하다.

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}, \quad \boldsymbol{R} = \boldsymbol{R}^{1/2}\boldsymbol{R}^{1/2} \tag{C.15}$$

D. 가능도비 검정

D.1. 가능도비 검정의 기초

가능도비 검정(likelihood ratio test)은 **제약 있는 모형**과 **제약 없는 모형**의 최대가능도 함수(maximum likelihood function)의 비율를 이용하여 두 모형을 비교하는 검정이며 통계적 가설 검정에 널리 사용되고 있다.

먼저 확률 변수 또는 확률 벡터 $\boldsymbol{X}$ 가 확률 밀도 함수 $f(\boldsymbol{X}|\boldsymbol{\theta})$ 를 따른다고 가정하자. 또한 다음과 같은 귀무 가설 검정을 고려하자. 또한 모수벡터 $\boldsymbol{\theta}$ 는 전체 모수 공간 $\boldsymbol{\Theta}$ 에 속한다고 가정한다.

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \text{vs} \quad H_a : \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0 \tag{D.1}$$

위의 가설에서 귀무 가설 H_0 는 모수 공간 $\boldsymbol{\Theta}$ 의 부분 집합 $\boldsymbol{\Theta}_0$ 에 모수가 속한다는 것 (모수에 대한 제약조건)을 의미한다.

이러한 가설 검정을 위하여 표본 벡터 $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ 에 대한 가능도 함수 $L(\boldsymbol{\theta})$ 와 로그 가능도 함수 $\ell(\boldsymbol{\theta})$는 다음과 같이 정의된다. 참고로 가능도 함수는 주어진 표본의 값에 대하여 모수 $\boldsymbol{\theta}$ 의 함수로 생각할 수 있다.

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(\boldsymbol{X}_i|\boldsymbol{\theta}) \\ \ell(\boldsymbol{\theta}) &= \log \prod_{i=1}^{n} f(\boldsymbol{X}_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log f(\boldsymbol{X}_i|\boldsymbol{\theta}) \end{aligned}$$

이제 최대 가능도 추정을 다음 두 개의 경우에 대하여 고려해 보자.

1. 제약이 있는 경우, 즉 귀무 가설이 참인 경우: $\theta \in \boldsymbol{\Theta}_0$

제약이 있는 경우에 대한 최대 가능도 추정량은 다음의 조건을 만족하는 경우

$$\hat{\boldsymbol{\theta}}_0 = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})$$

2. 제약이 없는 경우, 즉 귀무 가설이 거짓인 경우

제약이 없는 경우에 대한 최대 가능도 추정량은 다음의 조건을 만족하는 경우니다.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$$

이제 두 경우에 대한 최대 가능도 추정량을 이용하여 가능도비 검정 통계량 Λ 를 다음과 같이 정의한다.

$$\begin{aligned}\Lambda &= \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})} \\ &= \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})} \in (0,1]\end{aligned}$$

가능도비 Λ 는 귀무 가설이 참일 때 1에 가까운 값을 가지며, 실제 모수가 귀무 가설의 제약조건에서 멀어지면 0에 가까운 값을 가진다. 따라서 귀무 가설을 기각하기 위한 기각역은 Λ 가 작은 값이 되는 영역으로 설정한다. 또한 가설 검정의 편의성을 위하여 가능도비에 로그를 취하고 -2 를 취한 값을 검정에 이용한다.

$$\lambda = -2\log\Lambda = -2\left\{\ell(\hat{\boldsymbol{\theta}}_0) - \ell(\hat{\boldsymbol{\theta}})\right\} \in [0,\infty) \tag{D.2}$$

이제 위의 식에서 정의된 λ 는 값이 크면 클수록 귀무가설에 반대되는 증거이다. 따라서 λ의 값이 주어진 기각역 c 보다 크면 귀무 가설을 기각한다.

$$\text{reject } H_0 \quad \text{if } \lambda > c$$

기각역 c 는 λ 에 비례하는 적절한 검정 통계량을 찾은 다음, 주어진 확률 분포와 표본의 갯수에 따라서 검정 통계량의 정확한 분포를 구하여 유도할 수 있다. 하지만 대부분의 경우에는 λ 의 다음과 같은 점근적 성질(표본의 개수가 증가할 때 극한 분포를 이용)을 이용하여 기각역을 유도한다(Wilks' theorem)

$$\lambda = -2\log\Lambda \to_d \chi^2_\nu$$

위의 성질에서 $\to_d$ 는 분포의 수렴을 의미하며, χ^2_ν 는 자유도 ν 인 카이제곱 분포를 나타낸다. 자유도 ν 는 전체 모수 공간과 제약조건 공간 차원의 차이이며 다음과 같이 계산된다.

$$\nu = \dim(\boldsymbol{\Theta}) - \dim(\boldsymbol{\Theta}_0)$$

D.2. 다변량 정규분포의 가능도비 검정

D.2.1. 두 평균벡터의 비교

확률 벡터 $\boldsymbol{X}$ 과 $\boldsymbol{Y}$ 가 평균이 각각 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 p-차원 다변량 정규 분포를 따른다고 가정하자.

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \boldsymbol{Y} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

다변량 정규 분포의 확률밀도함수는 다음과 같이 주어진다.

$$f_p(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\}$$

더 나아가 다음과 같은 가설 검정을 고려하자.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs } H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

이제 가설 검정을 위하여 두 그룹에서 각각 n_1, n_2개의 다변량 표본이 관측되었다고 하자.

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_{n_1} \sim_{IID} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_{n_2} \sim_{IID} N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

로그 가능도 함수

이제 로그 가능도 함수를 정의하자. 먼저 제약 조건이 없는 경우를 고려하자.

$$\begin{aligned}
\ell(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= \log L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\
&= \log \prod_{i=1}^{n_1} f_p(\boldsymbol{X}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \prod_{i=1}^{n_2} f_p(\boldsymbol{Y}_i \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\
&= \sum_{i=1}^{n_1} \log f_p(\boldsymbol{X}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \sum_{i=1}^{n_2} \log f_p(\boldsymbol{Y}_i \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \\
&= -\frac{n_1+n_2}{2} \log |2\pi\boldsymbol{\Sigma}| \\
&\quad - \frac{1}{2}\left[\sum_{i=1}^{n_1} (\boldsymbol{X}_i - \boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_1) + \sum_{i=1}^{n_2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_2) \right]
\end{aligned} \tag{D.3}$$

만약 귀무 가설이 참이라면 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ 이므로 로그 가능도 함수는 다음과 같이 주어진다.

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \log \prod_{i=1}^{n_1} f_p(\boldsymbol{X}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{n_2} f_p(\boldsymbol{Y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \sum_{i=1}^{n_1} \log f_p(\boldsymbol{X}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{i=1}^{n_2} \log f_p(\boldsymbol{Y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= -\frac{n_1+n_2}{2} \log |2\pi\boldsymbol{\Sigma}| \\
&\quad - \frac{1}{2}\left[\sum_{i=1}^{n_1} (\boldsymbol{X}_i - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}) + \sum_{i=1}^{n_2} (\boldsymbol{Y}_i - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}) \right]
\end{aligned} \tag{D.4}$$

D.2.2. 재곱합의 분해

이제 이차형식의 다음과 같은 대각합(trace) 표현을 이용하면

$$\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} = tr(\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^t)$$

로그 가능도 함수에 나타나는 제곱합 항들을 다음과 같이 표현할 수 있다.

$$\begin{aligned}
\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1) &= \mathsf{tr}\left[\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)\right] \\
&= \sum_{i=1}^{n_1}\mathsf{tr}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)\right] \\
&= \sum_{i=1}^{n_1}\mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\right] \\
&= \mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\right]
\end{aligned} \tag{D.5}$$

또한 위의 식에서 $\boldsymbol{X}_i-\boldsymbol{\mu}_1$ 를 $(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)+(\bar{\boldsymbol{X}}_1-\boldsymbol{\mu}_1)$ 로 전개하면 평균 분해를 이용하여 다음과 같이 쓸 수 있다.

$$\begin{aligned}
\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t &= \sum_{i=1}^{n_1}\left[(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)+(\bar{\boldsymbol{X}}_1-\boldsymbol{\mu}_1)\right]\left[(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)+(\bar{\boldsymbol{X}}_1-\boldsymbol{\mu}_1)\right]^t \\
&= \sum_{i=1}^{n_1}(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)^t + n_1(\bar{\boldsymbol{X}}_1-\boldsymbol{\mu}_1)(\bar{\boldsymbol{X}}_1-\boldsymbol{\mu}_1)^t \\
&= n_1 S_x + n_1(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t
\end{aligned} \tag{D.6}$$

위의 식에서 S_x 는 확률 표본 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{n_1}$ 으로 만들어진 표본 공분산 행렬이다 (아래 식에서 공분산행렬의 추정에서 분포를 최대가능도 추정량으로 하여 n_1-1 대신 n_1 을 적용하였다)

$$S_X = \frac{1}{n_1}\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)^t$$

이제 식 D.6 을 식 D.5 에 적용하면 다음과 같이 쓸 수 있다.

$$
\begin{aligned}
\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1) &= \mathsf{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\right] \\
&= \mathsf{tr}\left[n_1\boldsymbol{\Sigma}^{-1}S_x + n_1\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t\right] \\
&= n_1\left[\mathsf{tr}(\boldsymbol{\Sigma}^{-1}S_x) + \mathsf{tr}(\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t)\right] \\
&= n_1\,\mathsf{tr}(\boldsymbol{\Sigma}^{-1}S_x) + n_1(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)
\end{aligned}
\tag{D.7}
$$

D.2.3. 로그 가능도 함수의 재표현

이제 분해식 식 D.7 를 식 D.3 에 적용하여 제약조건이 없는 경우의 로그 가능도 함수를 다음과 같이 표현할 수 있다.

$$
\begin{aligned}
\ell(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}) &= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}\left[\sum_{i=1}^{n_1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu}_1) + \sum_{i=1}^{n_2}(\boldsymbol{Y}_i-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_2)\right] \\
&= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}\Bigg[n_1\,\mathsf{tr}(\boldsymbol{\Sigma}^{-1}S_x) + n_1(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1) \\
&\qquad\qquad + n_2\,\mathsf{tr}(\boldsymbol{\Sigma}^{-1}S_y) + n_2(\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2)\Bigg] \\
&= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\,\mathsf{tr}(\boldsymbol{\Sigma}^{-1}[n_1S_x+n_2S_y]) \\
&\quad -\frac{1}{2}\left[n_1(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1) + n_2(\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2)\right] \\
&= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\,\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) \\
&\quad -\frac{1}{2}\left[n_1(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1) + n_2(\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2)\right]
\end{aligned}
\tag{D.8}
$$

위의 식에서 $\boldsymbol{W}$ 는 그룹내의 변동을 표시하는 제곱합 행렬이다. 참고로 식 3.4 에서 정의된 풀링된 공분산 행렬 $\boldsymbol{S}_p$ 와 다음과 같은 관계가 있다.

$$
\begin{aligned}
\boldsymbol{W} &= n_1S_X + n_2S_Y \\
&= \sum_{i=1}^{n_1}(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)(\boldsymbol{X}_i-\bar{\boldsymbol{X}}_1)^t + \sum_{i=1}^{n_2}(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}}_2)(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}}_2)^t \\
&= (n_1+n_2-2)\boldsymbol{S}_p
\end{aligned}
\tag{D.9}
$$

제약조건이 없는 가능도 함수 식 D.8 에 대해서 최대 가능도 추정을 적용하면 다음과 같은 평균에 대한 최대가능도 추정량은 각각 그룹의 표본 평균이 된다.

$$\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{X}}_1, \quad \hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{Y}}_1$$

이제 위의 평균에 대한 최대가능도 추정량을 제약 조건이 없는 가능도 함수 식 D.8 에 대입하면 다음과 같은 식을 얻게 된다.

$$\begin{aligned}\ell(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}) &= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) \\ &\quad - \frac{1}{2}\left[n_1(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}}_1)^t\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}}_1) + n_2(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}}_2)^t\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}}_2)\right] \\ &= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) + 0\end{aligned}$$

위의 식에서 공분산 행렬에 대한 최대가능도 추정량을 구하면 다음과 같은 추정량을 얻게되며

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n_1+n_2}\boldsymbol{W}$$

따라서 공분산 행렬에 대한 추정량을 제약조건이 없는 로그 가능도 함수에 대입하면 다음의 값을 얻게된다.

$$\begin{aligned}\ell(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}) &= -\frac{n_1+n_2}{2}\log|2\pi\hat{\boldsymbol{\Sigma}}| - \frac{1}{2}\mathsf{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}) \\ &= -\frac{n_1+n_2}{2}\log|2\pi\hat{\boldsymbol{\Sigma}}| - \frac{1}{2}\mathsf{tr}((n_1+n_2)\boldsymbol{W}^{-1}\boldsymbol{W}) \\ &= -\frac{n_1+n_2}{2}\log|2\pi\hat{\boldsymbol{\Sigma}}| - \frac{p(n_1+n_2)}{2}\end{aligned} \tag{D.10}$$

이제 제약조건 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ 가 있는 경우의 로그 가능도 함수를 고려하자. 식 D.8 의 마지막 항을 이용하면 다음과 같이 제약 조건이 있는 가능도 함수 식 D.4 를 다음과 같이 표현할 수 있다.

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\mathsf{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) \\ &\quad - \frac{1}{2}\left[n_1(\bar{\boldsymbol{X}}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) + n_2(\bar{\boldsymbol{Y}}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{Y}}-\boldsymbol{\mu})\right]\end{aligned} \tag{D.11}$$

위의 제약조건이 있는 로그 가능도 함수에 대하여 평균 벡터 $\boldsymbol{\mu}$ 에 최대 가능도 추정량을 구하면 다음과 같다.

$$\hat{\boldsymbol{\mu}} = \frac{n_1\bar{\boldsymbol{X}} + n_2\bar{\boldsymbol{Y}}}{n_1+n_2}$$

여기서 그룹 간의 변동을 나타내는 제곱합 행렬 $\boldsymbol{B}$ 는 다음과 같이 정의한다.

$$\boldsymbol{B} \equiv n_1(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}})(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}})^t + n_2(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}})(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}})^t \tag{D.12}$$

이제 평균 벡터의 추정량 $\hat{\boldsymbol{\mu}}$ 를 식 D.11 을 에 대입하면 다음과 같이 로그 가능도 함수가 나타나며

$$
\begin{aligned}
\ell(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) &= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) \\
&\quad - \frac{1}{2}\left[n_1(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}})^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}}) + n_2(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}})^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}})\right] \\
&= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}) \\
&\quad - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\{n_1(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}})(\bar{\boldsymbol{X}}-\hat{\boldsymbol{\mu}})^t + n_2(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}})(\bar{\boldsymbol{Y}}-\hat{\boldsymbol{\mu}})^t\}\right] \\
&= -\frac{n_1+n_2}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{W}+\boldsymbol{B}))
\end{aligned}
\tag{D.13}
$$

이제 오그 가능도 함수 식 D.13 에서 공분산 행렬 $\boldsymbol{\Sigma}$ 에 대한 최대 가능도 추정량을 구하면 다음과 같다.

$$
\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n_1+n_2}(\boldsymbol{W}+\boldsymbol{B})
$$

위의 공분산 행렬에 대한 추정량을 식 D.13 에 대입하면 제약조건이 있는 경우 로그 가능도 함수의 최대값은 다음과 같이 얻어진다.

$$
\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}_0) = -\frac{n_1+n_2}{2}\log|2\pi\hat{\boldsymbol{\Sigma}}_0| - \frac{p(n_1+n_2)}{2} \tag{D.14}
$$

D.2.4. 가능도비 검정 통계량

이제 로그가능도비 통계량 λ 를 식 D.2 에 정의된 식을 이용하여 다음과 같이 쓸 수 있다.

$$
\begin{aligned}
\lambda &= -2\log\Lambda \\
&= -2\left\{\ell\left(\hat{\mu},\, \hat{\Sigma}_0\right) - \ell\left(\hat{\mu}_1,\, \hat{\mu}_2,\, \hat{\Sigma}\right)\right\} \\
&= N\log\left|\hat{\Sigma}_0\right| - N\log\left|\hat{\Sigma}\right| \\
&= N\log\left(\frac{\left|\hat{\Sigma}_0\right|}{\left|\hat{\Sigma}\right|}\right) \\
&= N\log\left(\frac{|W+B|}{|W|}\right),
\end{aligned}
\tag{D.15}
$$

위의 식에서 $N = n_1 + n_2$ 이다.

마지막으로 두 집단에서는 다음과 같은 행렬식 공식를 이용하자 (부록의 Section A.9 참조)

$$
|\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^t| = |\boldsymbol{A}|(\boldsymbol{1}+\boldsymbol{v}^t\boldsymbol{A}^{-1}\boldsymbol{u}) \tag{D.16}
$$

위의 정리를 이용하기 위하여 식 D.12 정의된 그룹간 제곱합 행렬 $\boldsymbol{B}$ 를 다음과 같이 표현해 보자

$$
\begin{aligned}
\boldsymbol{B} &= n_1(\bar{\boldsymbol{X}} - \hat{\boldsymbol{\mu}})(\bar{\boldsymbol{X}} - \hat{\boldsymbol{\mu}})^t + n_2(\bar{\boldsymbol{Y}} - \hat{\boldsymbol{\mu}})(\bar{\boldsymbol{Y}} - \hat{\boldsymbol{\mu}})^t \\
&= n_1 \left[\bar{\boldsymbol{X}} - \frac{n_1\bar{\boldsymbol{X}} + n_2\bar{\boldsymbol{Y}}}{N}\right] + n_2 \left[\bar{\boldsymbol{Y}} - \frac{n_1\bar{\boldsymbol{X}} + n_2\bar{\boldsymbol{Y}}}{N}\right]^t \\
&= \frac{n_1 n_2}{N}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^t \\
&= \alpha \boldsymbol{d}\boldsymbol{d}^t,
\end{aligned}
\tag{D.17}
$$

위의 식에서

$$
\alpha = \frac{n_1 n_2}{N}, \quad d = \bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}
$$

따라서 식 D.16 에서 $\boldsymbol{u} = \boldsymbol{v} = \sqrt{\alpha}\boldsymbol{d}$ 로 놓으면

$$
\begin{aligned}
|\boldsymbol{W} + \boldsymbol{B}| &= |\boldsymbol{W} + \alpha \boldsymbol{d}\boldsymbol{d}^t| \\
&= |\boldsymbol{W}| \left(\mathbf{1} + \alpha \boldsymbol{d}^t \boldsymbol{W}^{-1} \boldsymbol{d}\right)
\end{aligned}
$$

이제 위의 식을 식 D.15 에 넣고 정리하면 다음과 같은 결과를 얻는다.

$$
\begin{aligned}
\lambda &= -2 \log \Lambda \\
&= N \log \left(\frac{|W + B|}{|W|}\right) \\
&= N \log \left(\frac{|\boldsymbol{W}| \left(1 + \alpha \boldsymbol{d}^t \boldsymbol{W}^{-1} \boldsymbol{d}\right)}{|\boldsymbol{W}|}\right) \\
&= N \log \left(1 + \alpha \, d^\top W^{-1} d\right) \\
&= N \log \left(1 + \frac{T^2}{N-2}\right)
\end{aligned}
\tag{D.18}
$$

위의 식에서 주어진 T^2 는 식 3.5 에서 장의한 Hotelling의 T^2 통계량이다. 따라서 로그 가능도비 검정 통계량 λ 와 Hotelling의 T^2 통계량은 단조 증가 함수 관계에 있음을 알 수 있다. 이러한 결과로서 Hotelling의 T^2 을 이용한 검정은 가능도비 검정이다.