STAT 3690 Lecture Note

Week Three (Jan 23, 25, & 27, 2023)

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2023/Jan/27 10:59:22

Statistical modelling (con'd)

Transformation of random vectors

- Derive the pdf of continuous $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ from the pdf of continuous $\boldsymbol{X}$
- Prerequisite
  - $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_p]^\top$
  - $\boldsymbol{g} = (g_1, \ldots, g_p)$: $\mathbb{R}^p \to \mathbb{R}^p$ is a continuous one-to-one map with inverse $\boldsymbol{g}^{-1} = (h_1, \ldots, h_p)$, i.e., $Y_i = g_i(\boldsymbol{X})$ and $X_i = h_i(\boldsymbol{Y})$
- Elaborate $\mathrm{supp}(\boldsymbol{Y}) = \{[y_1, \ldots, y_p]^\top : [h_1(y_1, \ldots, y_p), \ldots, h_p(y_1, \ldots, y_p)]^\top \in \mathrm{supp}(\boldsymbol{X})\}$
- Jacobian matrix of $\boldsymbol{g}^{-1}$ is $\mathbf{J}_{\boldsymbol{g}^{-1}} = [\partial x_i/\partial y_j]_{p\times p} = [\partial h_i(y_1, \ldots, y_p)/\partial y_j]_{p\times p}$
  - Also, $|\det(\mathbf{J}_{\boldsymbol{g}^{-1}})| = |\det([\partial y_i/\partial x_j]_{p\times p})|^{-1} = |\det([\partial g_i(x_1, \ldots, x_p)/\partial x_j]_{p\times p})|^{-1}$
- Then

$$f_{\boldsymbol{Y}}(y_1, \ldots, y_p) = f_{\boldsymbol{X}}(h_1(y_1, \ldots, y_p), \ldots, h_p(y_1, \ldots, y_p))|\det(\mathbf{J}_{\boldsymbol{g}^{-1}})|\mathbf{1}_{\mathrm{supp}(\boldsymbol{Y})}(y_1, \ldots, y_p)$$

- Exercise: Let $\boldsymbol{X} = [X_1, X_2]^\top$ follow the standard bivariate normal, i.e., its pdf is

$$f_{\boldsymbol{X}}(x_1, x_2) = (2\pi)^{-1}\exp\{-(x_1^2 + x_2^2)/2\}\mathbf{1}_{\mathbb{R}^2}(x_1, x_2).$$

  Find out the joint pdf of $\boldsymbol{Y} = [Y_1, Y_2]^\top$, where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \le Y_2 < 2\pi$ is the angle from the positive x-axis to the ray from the origin to the point (X_1, X_2), that is, Y is X in the polar coordinate.

- Exercise: Given positive α, β and θ, $\boldsymbol{X} = [X_1, X_2]^\top$ follow

$$f_{\boldsymbol{X}}(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} x_1^{\alpha-1} x_2^{\beta-1} \exp\left(-\frac{x_1 + x_2}{\theta}\right)\mathbf{1}_{\mathbb{R}^+\times\mathbb{R}^+}(x_1, x_2),$$

  where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$, e.g., $\Gamma(n) = (n-1)!$ for positive integer n. Find out the joint pdf of $\boldsymbol{Y} = [Y_1, Y_2]^\top$, where $Y_1 = X_1/(X_1 + X_2)$ and $Y_2 = X_1 + X_2$.

Mean matrix

- $\mathrm{E}(\boldsymbol{X}) = [\mathrm{E}(X_{ij})]_{n\times p}$, where
  - Random $n \times p$ matrix $\boldsymbol{X} = [X_{ij}]_{n\times p}$
- (Linearity) $\mathrm{E}(\mathbf{A}\boldsymbol{X} + \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\mathrm{E}(\boldsymbol{X}) + \mathbf{B}\mathrm{E}(\boldsymbol{Y})$, where
  - Fixed $\mathbf{A} \in \mathbb{R}^{\ell\times n}$ and $\mathbf{B} \in \mathbb{R}^{\ell\times m}$
  - Random matrices $\boldsymbol{X} = [X_{ij}]_{n\times p}$ and $\boldsymbol{Y} = [Y_{ij}]_{m\times p}$

Covariance matrix

- Random p-vector $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ and random q-vector $\boldsymbol{Y} = [Y_1, \ldots, Y_q]^\top$
- Covariance matrix (defined via expectation) $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{E}[\{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\}\{\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})\}^\top]$
  - Also, $\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \mathrm{E}(\boldsymbol{X}\boldsymbol{Y}^\top) - \mathrm{E}(\boldsymbol{X})\mathrm{E}(\boldsymbol{Y}^\top)$
  - The (i, j)-entry of $\boldsymbol{\Sigma}_{\boldsymbol{XY}}$ is $\mathrm{cov}(X_i, Y_j)$
- $\boldsymbol{\Sigma}_{\mathbf{A}\boldsymbol{X}+\boldsymbol{a},\mathbf{B}\boldsymbol{Y}+\boldsymbol{b}} = \mathbf{A}\Sigma_{\boldsymbol{XY}}\mathbf{B}^\top$ for fixed $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\boldsymbol{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{\ell\times q}$ and $\boldsymbol{b} \in \mathbb{R}^\ell$
- $\boldsymbol{\Sigma}_{\boldsymbol{X}} \geq 0$, where $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathrm{cov}(\boldsymbol{X})$ is short for $\boldsymbol{\Sigma}_{\boldsymbol{XX}} = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})$

Sample covariance matrix

- Samples $\boldsymbol{X}_k = [X_{k1}, \ldots, X_{kp}]^\top$ and $\boldsymbol{Y}_k = [Y_{k1}, \ldots, Y_{kq}]^\top$, $k = 1, \ldots, n$
- $(\boldsymbol{X}_k, \boldsymbol{Y}_k) \overset{\text{iid}}{\sim} (\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X} = [X_1, \ldots, X_p]^\top$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_q]^\top$
- Sample mean vectors
  - $\bar{\boldsymbol{X}} = n^{-1}\sum_{k=1}^n \boldsymbol{X}_k = [\bar{X}_{\cdot 1}, \cdots, \bar{X}_{\cdot p}]^\top$
    - $\bar{X}_{\cdot i} = n^{-1}\sum_{k=1}^n X_{ki}$
  - $\bar{\boldsymbol{Y}} = n^{-1}\sum_{k=1}^n \boldsymbol{Y}_k = [\bar{Y}_{\cdot 1}, \cdots, \bar{Y}_{\cdot q}]^\top$
    - $\bar{Y}_{\cdot j} = n^{-1}\sum_{k=1}^n Y_{kj}$
- Sample covariance matrix:

$$\mathbf{S}_{\boldsymbol{XY}} = \frac{1}{n-1}\sum_{k=1}^n \{(\boldsymbol{X}_k - \bar{\boldsymbol{X}})(\boldsymbol{Y}_k - \bar{\boldsymbol{Y}})^\top\}$$

  - The (i, j)-entry of $\mathbf{S}_{\boldsymbol{XY}}$ is $(n-1)^{-1}\sum_{k=1}^n (X_{ki} - \bar{X}_{\cdot i})(Y_{kj} - \bar{Y}_{\cdot j})$, i.e., the sample covariance between X_i (the ith entry of $\boldsymbol{X}$) and Y_j (the jth entry of $\boldsymbol{Y}$)
  - Unbiasedness: $\mathrm{E}(\mathbf{S}_{\boldsymbol{XY}}) = \boldsymbol{\Sigma}_{\boldsymbol{XY}}$
  - $\mathbf{S}_{\mathbf{A}\boldsymbol{X}+\boldsymbol{a},\mathbf{B}\boldsymbol{Y}+\boldsymbol{b}} = \mathbf{A}\mathbf{S}_{\boldsymbol{XY}}\mathbf{B}^\top$ for $\mathbf{A} \in \mathbb{R}^{m\times p}$, $\boldsymbol{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{\ell\times q}$ and $\boldsymbol{b} \in \mathbb{R}^\ell$
  - $\mathbf{S}_{\boldsymbol{X}} \geq 0$
  - Implementation in *R*: cov() (or var() if $\boldsymbol{X} = \boldsymbol{Y}$)

Computing sample mean vectors and sample covariance matrices via *R*

```
options(digits = 4)
set.seed(1)

# mean vector and covariance matrix
(Mu = runif(3))
```

```r
A = matrix(runif(15), nrow = 3, ncol = 5)
(Sigma = A %*% t(A))

# generation of samples
n = 100
sample = MASS::mvrnorm(n, Mu, Sigma)
colnames(sample) = c('W', 'H', "BP")
head(sample)

# reference for various scatterplots https://www.statmethods.net/graphs/scatterplot.html
# scatterplots for paired features
pairs(sample)
# (spinning) 3D scatterplot
rgl::plot3d(sample[,1], sample[,2], sample[,3], col = "red", size = 6)

# sample mean vector for [V1,V2,V3]^T
(MuHat = apply(sample, 2, mean))
(MuHat = colMeans(sample))
# sample covariance matrix for [W,H,BP]^T
## following the definition
S = 0; for (i in 1:n){S = S + 1/(n-1) * (sample[i,]-MuHat) %*% t(sample[i,]-MuHat)}; S
## via var()
(S = var(sample))
## via cov()
(S = cov(sample))

var(sample[,2], sample[,1])

# sample covariance matrix for W & [H,BP]^T
cov(sample[,1], sample[,2:3])
# sample covariance matrix for H & [BP,W]^T
cov(sample[,2], sample[,c(3,1)])

# another sample
(Mu2 = runif(2))
A2 = matrix(runif(10), nrow = 2, ncol = 5)
(Sigma2 = A2 %*% t(A2))
sample2 = MASS::mvrnorm(n, Mu2, Sigma2)
colnames(sample2) = c('CH', 'HR')
head(sample2)
cov(sample, sample2)
sample_c = cbind(sample, sample2)
cov(sample_c)
cov(sample, sample2)
```

Multivariate normal (MVN) distribution (J&W Sec 4.2)

Definition

- Standard MVN
  - $\boldsymbol{Z} = [Z_1, \ldots, Z_p]^\top \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I}) \Leftrightarrow Z_1, \ldots, Z_p \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
  - pdf

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = (2\pi)^{-p/2} \exp(-\boldsymbol{z}^\top \boldsymbol{z}/2) \cdot \mathbf{1}_{\mathbb{R}^p}(\boldsymbol{z})$$

- General MVN
  - $\boldsymbol{X} = [X_1, \ldots, X_p]^\top \sim \mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Leftrightarrow$ there exists $\boldsymbol{\mu} \in \mathbb{R}^p$, $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{Z} \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I})$ such that $\boldsymbol{X} = \mathbf{A}\boldsymbol{Z} + \boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$
    - Limited to non-degenerate cases, i.e., invertible $\mathbf{A}$ ($\Leftrightarrow \boldsymbol{\Sigma} > 0$)
  - pdf

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-p/2}(\det\boldsymbol{\Sigma})^{-1/2}\exp\{-(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\}\cdot\mathbf{1}_{\mathbb{R}^p}(\boldsymbol{x})$$