STAT 3690 Lecture 07

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Feb 7, 2022

Useful properties of MVN

- $\mathbf{X} \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Leftrightarrow \mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim MVN_p(\mathbf{0}, \mathbf{I})$. So, we have a stochastic representation of arbitrary $\mathbf{X} \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} + \boldsymbol{\mu}$, where $\mathbf{Z} \sim MVN_p(\mathbf{0}, \mathbf{I})$.
- $\mathbf{X} \sim MVN$ iff, for all $a \in \mathbb{R}^p$, $a^\top \mathbf{X}$ has a (univariate) normal distribution.
- If $\mathbf{X} \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{X} + \boldsymbol{b} \sim MVN_q(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ for $\mathbf{A} \in \mathbb{R}^{q \times p}$ and $\text{rk}(\mathbf{A}) = q$.

- Exercise: Generate six iid samples following bivariate normal $MVN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = [3, 6]^\top, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 2 \\ 2 & 5 \end{bmatrix}.$$

```r
options(digits = 4)
set.seed(1)
Mu = matrix(c(3, 6), ncol = 1, nrow = 2)
Sigma = matrix(c(10, 2 ,2, 5), ncol = 2, nrow = 2)
n = 1000
# Method 1: via rnorm()
spectral = eigen(Sigma)
SigmaRoot = spectral$vectors %*% diag(spectral$values^.5) %*% t(spectral$vectors)
A1 = matrix(0, nrow = n, ncol = length(Mu))
for (i in 1:n) {
  A1[i, ] = t(SigmaRoot %*% matrix(rnorm(2), nrow = 2, ncol = 1) + Mu)
}
# Method 2: via MASS::mvrnorm()
A2 = MASS::mvrnorm(n, Mu, Sigma)
```

- Exercise:
  1. Prove that $(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(p)$ if $\mathbf{X} \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  2. Suppose $X_1 \sim N(0, 1)$. In the following two cases, verify that $X_2 \sim N(0, 1)$ as well. Does $\mathbf{X} = [X_1, X_2]^\top$ follow an MVN in both cases?
     a. $X_2 = -X_1$;
     b. $X_2 = (2Y - 1)X_1$, where $Y \sim Ber(p)$ and $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}$.
        - P.S.: $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \Leftrightarrow f_{\mathbf{Z}}(\boldsymbol{z}) = f_{\mathbf{X}}(\boldsymbol{x}) f_{\mathbf{Y}}(\boldsymbol{y})$, where $\mathbf{Z} = [\mathbf{X}^\top, \mathbf{Y}^\top]^\top$

- Solution:

1. $\mathbf{Y} = [Y_1, \dots, Y_p]^\top = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. Then $(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Y}^\top \mathbf{Y} = \sum_i Y_i^2 \sim \chi^2(p)$ since $Y_i \overset{\text{iid}}{\sim} N(0,1)$.
2. a. $\Pr(X_2 < x) = \Pr(X_1 > -x) = \Pr(X_1 < x)$ since pdf of $N(0,1)$ is symmetric with respect to x-axis; $\Pr(X_2 = -X_1) = 1 \Rightarrow \operatorname{supp}(\mathbf{X}) = \{(x, -x) \mid x \in \mathbb{R}\}$;
   b. $\Pr(X_2 < x) = \Pr(X_2 < x \mid Y = 0)\Pr(Y = 0) + \Pr(X_2 < x \mid Y = 1)\Pr(Y = 1) = (1 - p)\Pr(x_1 > -x) + p\Pr(X_1 < x) = \Pr(X_1 < x)$. $\Pr(X_2 = X_1) = \Pr(Y = 1) = p$ and $\Pr(X_2 = -X_1) = \Pr(Y = 0) = 1 - p \Rightarrow \operatorname{supp}(\mathbf{X}) = \{(x, -x) \mid x \in \mathbb{R}\} \bigcup \{(x, x) \mid x \in \mathbb{R}\}$.

```
options(digits = 4)
set.seed(1)
xsize = 1e4L
x1 = rnorm(xsize)
# case 1
x2 = -x1
plot3D::hist3D(z=table(cut(x1, 100), cut(x2, 100)), border = "black") # 3d histogram of (x1, x2)
plot3D::image2D(z=table(cut(x1, 100), cut(x2, 100)), border = "black") # plot the support of joint pdf
# case 2
Y = rbinom(n = xsize, 1, .3)
x2 = (2 * Y - 1) * x1
plot3D::hist3D(z=table(cut(x1, 100), cut(x2, 100)), border = "black") # 3d histogram of (x1, x2)
plot3D::image2D(z=table(cut(x1, 100), cut(x2, 100)), border = "black") # plot the support of joint pdf
```

Joint, marginal and conditional MVN

- If $\mathbf{X} \sim MVN_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ and

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

  with $\mathbf{\Sigma}_{11} > 0$ and $\mathbf{\Sigma}_{22} > 0$, then
  - $\mathbf{X}_i \sim MVN_{p_i}(\boldsymbol{\mu}_i, \mathbf{\Sigma}_{ii})$, i.e., marginals of MVN are MVN.
  - $\mathbf{X}_i \mid \mathbf{X}_j = \boldsymbol{x}_j \sim MVN_{p_i}(\boldsymbol{\mu}_{i|j}, \mathbf{\Sigma}_{i|j})$, i.e., conditionals of MVN are MVN.
  - $\boldsymbol{\mu}_{i|j} = \boldsymbol{\mu}_i + \mathbf{\Sigma}_{ij}\mathbf{\Sigma}_{jj}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_j)$
  - $\mathbf{\Sigma}_{i|j} = \mathbf{\Sigma}_{ii} - \mathbf{\Sigma}_{ij}\mathbf{\Sigma}_{jj}^{-1}\mathbf{\Sigma}_{ji}$
  - $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \Leftrightarrow \mathbf{\Sigma}_{12} = \mathbf{0}$

- Exercise: The argument $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \Leftrightarrow \mathbf{\Sigma}_{12} = 0$ is based on the assumption that $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top]^\top$ is of MVN. That is, if $\mathbf{X}_1$ and $\mathbf{X}_2$ are both MVN BUT they are not jointly normal, a zero $\mathbf{\Sigma}_{12}$ doesn't suffice for the independence between $\mathbf{X}_1$ and $\mathbf{X}_2$. A counter-example will be part of Assignment 1.

Checking normality (J&W Sec 4.6)

- Checking the univariate marginal distributions
  - Normal Q-Q plot
    - qqnorm(); car::qqPlot()
  - Normality test
    - shapiro.test()
- Checking the quadratic form
  - χ^2 Q-Q plot
    - $D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^\top \mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \approx \chi^2(p)$ if $\mathbf{X}_i \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}, \mathbf{\Sigma})$
    - qqplot(); car::qqPlot()

```
options(digits = 4)
install.packages(c("car"))
library(datasets)
data(iris)
head(iris)
dataframe = iris[,1:3]
p = ncol(dataframe)
n = nrow(dataframe)

# Marginal normal Q-Q plot
car::qqPlot(rnorm(n), id = F)
car::qqPlot(dataframe[,1], id = F)
car::qqPlot(dataframe[,2], id = F)

# Univariate normality test
shapiro.test(rnorm(n))
shapiro.test(dataframe[,1])
shapiro.test(dataframe[,2])

# chi^2 Q-Q plot
d_square = diag(
  as.matrix(sweep(dataframe, 2, colMeans(dataframe))) %*%
    solve(var(dataframe)) %*%
    t(as.matrix(sweep(dataframe, 2, colMeans(dataframe))))
)
car::qqPlot(d_square, dist="chisq", df = p, id = F)
```

Detecting outliers (J&W Sec 4.7)

- Scatter plot of standardized values
- Check the points farthest from the origin in χ^2 Q-Q plot

```
options(digits = 4)

# Scatter plot of standardized values
plot(1:n, scale(dataframe[,1]), ylab='Studentized', xlab='Label'); abline(2, 0); abline(-2, 0)
which(abs(scale(dataframe[,1]))>2)

# six points farthest from the origin in $\chi^2$ Q-Q plot
tail(order(d_square, decreasing=F))
```

Improving normality (J&W Sec 4.8)

- Box-Cox transformation: for $x > 0$,

$$x^*(\lambda) = \begin{cases} (x^\lambda - 1)/\lambda & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

  - If $x \leq 0$, change it to be positive first.

```
options(digits = 4)
install.packages(c("car","EnvStats"))

lambda = EnvStats::boxcox(dataframe[,1], optimize=T)$lambda
```

```
dataframe.new = (dataframe[,1]^lambda-1)/lambda

car::qqPlot(dataframe.new, id = F)
shapiro.test(dataframe.new)
```

- Exploratory data analysis (EDA)
    - J. Tukey (1977). Exploratory Data Analysis. Addison-Wesley. ISBN 978-0-201-07616-5.

R package “MVN”