STAT 3690 Lecture Note

Week Four (Jan 30, Feb 1, & 3, 2023)

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

2023/Feb/03 12:47:40

Multivariate normal (MVN) distribution (con'd, 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
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-p/2}(\det\boldsymbol{\Sigma})^{-1/2} \exp\{-(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\} \cdot \mathbf{1}_{\mathbb{R}^p}(\boldsymbol{x})$$

- Exercise: Density of $\mathrm{MVN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ evaluated at $(4, 7)$, where
$$\boldsymbol{\mu} = [3, 6]^\top, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 2 \\ 2 & 5 \end{bmatrix}.$$

```r
options(digits = 4)
(Mu = matrix(c(3, 6), ncol = 1, nrow = 2))
(Sigma = matrix(c(10, 2 ,2, 5), ncol = 2, nrow = 2))
(x = c(4,7))
# Method 1: following the pdf
(2*pi)^{-length(Mu)/2}*det(Sigma)^{-.5}*exp(-drop(t(x-Mu)%*%solve(Sigma)%*%(x-Mu))/2)
# Method 2: via mvtnorm::dmvnorm()
mvtnorm::dmvnorm(x, mean = Mu, sigma = Sigma)
```

Properties of MVN

- $\boldsymbol{X}$ is of MVN $\Leftrightarrow a^\top \boldsymbol{X}$ is normally distributed for ALL non-zero $a \in \mathbb{R}^p$.
  - Warning: the marginal normality do not imply the joint normality.
- If $\boldsymbol{X} \sim \mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}\boldsymbol{X}+\boldsymbol{b} \sim \mathrm{MVN}_q(\mathbf{A}\boldsymbol{\mu}+\boldsymbol{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ for $\mathbf{A} \in \mathbb{R}^{q\times p}$ of full-row-rank. Specifically, if $\boldsymbol{X} \sim \mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then
  - $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I})$ AND

- (Stochastic representation of MVN) there is $\boldsymbol{Z} \sim \text{MVN}_p(\mathbf{0}, \mathbf{I})$ such that $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}$.

- $(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2(p)$ if $\boldsymbol{X} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- Exercise: Generate six iid samples following bivariate normal $\text{MVN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 = 10
# Method 1: following the stochastic representation
sample1 = matrix(0, nrow = n, ncol = length(Mu))
for (i in 1:n) {
  sample1[i, ] = t(
    expm::sqrtm(Sigma) %*%
    matrix(rnorm(length(Mu)), nrow = length(Mu), ncol = 1) +
    Mu
 )
}
sample1
# Method 2: via MASS::mvrnorm()
(sample2 = MASS::mvrnorm(n, Mu, Sigma))
```

- Exercise: Suppose $X_1 \sim \mathcal{N}(0, 1)$. In the following two cases, verify that $X_2 \sim \mathcal{N}(0, 1)$ as well. Does $\boldsymbol{X} = [X_1, X_2]^\top$ follow an MVN in both cases?
  a. $X_2 = -X_1$;
  b. $X_2 = (2Y - 1)X_1$, where $Y \sim \text{Ber}(p)$ and $Y \perp\!\!\!\perp X_1$.

```r
options(digits = 4)
set.seed(1)
xsize = 1e4L
x1 = rnorm(xsize)
# case a
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 b
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
```

Marginal and conditional MVN

- If $\boldsymbol{X} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

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

  with
  - random p_i-vector $\boldsymbol{X}_i$, $i = 1, 2$,
  - p_i-vector $\boldsymbol{\mu}_i$, $i = 1, 2$,
  - $p_i \times p_i$ matrix $\boldsymbol{\Sigma}_{ii} > 0$, $i = 1, 2$,
- then
  - (Marginals of MVN are still MVN) $\boldsymbol{X}_i \sim \text{MVN}_{p_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$
  - $\boldsymbol{X}_i \mid \boldsymbol{X}_j = \boldsymbol{x}_j \sim \text{MVN}_{p_i}(\boldsymbol{\mu}_{i|j}, \boldsymbol{\Sigma}_{i|j})$
    - $\boldsymbol{\mu}_{i|j} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_j)$
    - $\boldsymbol{\Sigma}_{i|j} = \boldsymbol{\Sigma}_{ii} - \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{ji}$
  - $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 \Leftrightarrow \boldsymbol{\Sigma}_{12} = \boldsymbol{0}$
    - Warning: the prerequisite for this equivalence is the joint normal of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

- Exercise: The argument $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 \Leftrightarrow \boldsymbol{\Sigma}_{12} = 0$ is based on $[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top]^\top \sim \text{MVN}$. That is, if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are both MVN BUT they are not jointly normal, the zero $\boldsymbol{\Sigma}_{12}$ doesn't suffice for the independence between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Recall the case b. in the previous exercise: $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 = (2Y - 1)X_1$, where $Y \sim \text{Ber}(p)$ and $Y \perp\!\!\!\perp X_1$. Verify that X_1 and X_2 are not independent of each other. (Hint: assume the independence and then check the support of $[X_1, X_2]^\top$.)

Hypothesis testing

- Is it a squirrel?

Figure 1: Squirrel (Photograph by the Lacoste Garden Centre)

Figure 2: Flying Squirrel (Photograph by Joel Sartore)

Figure 3: Flying Squirrel (Photograph by Alex Badyaev)

- Null and alternative hypotheses, say H_0 and H_1, resp.
- Name of testing method
- Test statistic (varying with the testing method) and corresponding level α rejection region R_α
  - $\Pr(\text{test statistic} \in R_\alpha \mid H_0) \le \alpha$
  - Reject H_0 if the value of test statistic $\in R_\alpha$
    - Type I error: H_0 is incorrectly rejected; i.e., H_0 is correct but rejected

        * Type II error: H_0 is incorrectly accepted i.e., H_0 is wrong but NOT rejected

- p-value: a special test statistic with a default level α rejection region $[0, \alpha]$
- Necessary components in reporting a testing result
  1. Hypotheses
  2. Name of approach
  3. Level α
  4. (Value of test statistic AND rejection region) OR p-value
  5. Conclusion: e.g., at the α level, we reject/do not reject H_0, i.e., we believe that...

Checking/testing the normality (J&W Sec 4.6)

- Checkcing the univariate marginal distributions
  - Normal Q-Q plot
    * qqnorm(); car::qqPlot()
  - Univariate normality test
    * shapiro.test(); nortest::ad.test(); MVN::mvn()
- Checkcing the multivariate normality
  - χ^2 Q-Q plot
    * $D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \mathbf{S}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \approx \chi^2(p)$ if $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
    * qqplot(); car::qqPlot()
  - Multivariate normality test
    * MVN::mvn()

```r
options(digits = 4)
library(datasets)
data(iris)
head(iris)
(iris_setosa = iris[iris$Species=='setosa', 1:3])
p = ncol(iris_setosa)
n = nrow(iris_setosa)

# Marginal normal Q-Q plot
car::qqPlot(rnorm(n), id = F)
car::qqPlot(iris_setosa[,1], id = F)
car::qqPlot(iris_setosa[,2], id = F)
car::qqPlot(iris_setosa[,3], id = F)

# Univariate normality test
## Shapiro-Wilk Normality Test
shapiro.test(rnorm(n))
shapiro.test(iris_setosa[,1])
shapiro.test(iris_setosa[,2])
shapiro.test(iris_setosa[,3])
## Anderson-Darling test for normality
nortest::ad.test(iris_setosa[,1])
nortest::ad.test(iris_setosa[,2])
nortest::ad.test(iris_setosa[,3])
MVN::mvn(
  iris_setosa,
  univariateTest = "AD" # "SW"/"CVM"/"Lillie"/"SF"/"AD"
)$univariateNormality
```

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

# Multivariate normality test
MVN::mvn(
  iris_setosa,
  mvnTest = "dh" # "mardia"/"hz"/"royston"/"dh"/"energy"
)$multivariateNormality
```