Multivariate Analysis Lecture 4: A Random Sample from A Multivariate Distribution

Zhaoxia Yu
Professor, Department of Statistics

2023-04-13

Section 1

Review of Lecture 03

Outline

- Definitions
- A random variable from a univariate distribution
- A random sample from a univariate distribution
- A random vector from a multivariate distribution
- A random sample from a multivariate distribution
- Linear combinations of a random vector

Definitions: Mean and Variance

- Let X be a random variance. Its mean, denoted by $E[X]$, is defined as

$$\mu = E[X] = \begin{cases} \int_{-\infty}^{\infty} xf(x)dx & \text{if } X \text{ is continuous} \\ \sum_i x_i p_i & \text{if } X \text{ is discrete} \end{cases}$$

- Its variance, denoted by $Var(X)$, is defined as $E[(X-\mu)^2]$,

$$\sigma^2 = Var[X] = E[(X-\mu)^2] = \begin{cases} \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \\ \sum_i (x_i-\mu)^2 p_i & \text{if } X \text{ is discrete} \end{cases}$$

Definitions: Mean Vector and Covariance Matrix

- Let $\mathbf{X}_{p\times 1} = (X_1, \cdots, X_p)^T$ be a random vector. Its mean is defined as

$$E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_p] \end{pmatrix}$$

- Its covariance matrix is defined as

$$\mathbf{\Sigma}_{p\times p} = Cov(\mathbf{X}) = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$$

Random Vectors: An Example

- Let $X_1, \cdots, X_n$ be iid random variables from a univariate distribution with mean μ and variance σ^2. We say $X_1, \cdots, X_n$ form a random sample.
- We often stack the random variables vertically:

$$\mathbf{X}_{n\times 1} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},$$

- Remark: $\mathbf{X}$ is a random vector with mean vector and covariance matrix

$$E[\mathbf{X}] = \mu \mathbf{1}_n = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, Cov(\mathbf{X}) = \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

A Random Vector From a Multivariate Distribution

- Let $\mathbf{X}_{p\times 1}$ be a random vector from a multivariate distribution with mean vector $\boldsymbol{\mu}_{p\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_{p\times p}$.
- In other words,
  - $E(\mathbf{X}) = \boldsymbol{\mu}$
  - $Cov(\mathbf{X}) = \boldsymbol{\Sigma}$.

A Random Sample From a Multivariate Distribution

- Consider a random sample $\mathbf{X}_1, \cdots, \mathbf{X}_n$ from a multivariate distribution with mean vector $\boldsymbol{\mu}_{p\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_{p\times p}$.
- We often stack the random vectors to form an $n \times p$ matrix:

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix}$$

A Random Sample From a Multivariate Distribution: Sample Mean Vector

- Sample mean vector is

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i = (\frac{1}{n}\mathbf{1}_n^T\mathbf{X})^T$$

It is a random vector with

- mean vector $E[\bar{\mathbf{X}}] = \boldsymbol{\mu}$, i.e., the sample mean vector is unbaised for the population mean vector. $\bar{\mathbf{X}}$ can be used to estimate $\boldsymbol{\mu}$.
- covariance matrix $Cov(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$

A Random Sample From a Multivariate Distribution: Sample Covariance Matrix

- The sample covariance matrix is

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$$

- It is unbiased for $\boldsymbol{\Sigma}$, i.e., $E[\mathbf{S}] = \boldsymbol{\Sigma}$.
- We showed that

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{C}\mathbf{X}$$

  where $\mathbf{C}_{n\times n} = \mathbf{I} - \frac{1}{n}\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$
- This expression is helpful when we derive the distribution of $\mathbf{S}$.

Section 2

Linear Combination of a Random Vector

Definition of a Linear Combination of a Random Vector

- Let $\mathbf{X}_{p\times 1} = (X_1, \cdots, X_n)^T$ be a p-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- Consider a linear combination of the form:

$$Y = \mathbf{a}^T\mathbf{X} = \sum_{i=1}^{p} a_i X_i$$

where $\mathbf{a}$ is a p-dimensional constant vector.

- Note $Y = \mathbf{a}^T\mathbf{X} = \mathbf{X}^T\mathbf{a}$, both gives the same univariate random variable Y.
- E.g., $\mathbf{X} = (X_1, X_2, X_3)^T$, $a = (1/3, 1/3, 1/3)^T$. Then

$$Y = a^T X = \frac{1}{3}(X_1 + X_2 + X_3)$$

Mean of $Y = a^T X$

- The mean of Y can be expressed as:

$$\begin{aligned} E(Y) &= E(\mathbf{a}^T \mathbf{X}) \\ &= \mathbf{a}^T E(\mathbf{X}) \\ &= \mathbf{a}^T \boldsymbol{\mu} \end{aligned}$$

- Intuitively, the mean of Y is a weighted average of the components of $\mathbf{X}$, with weights given by the corresponding components of $\mathbf{a}$.

Variance of Y

- $Var(Y) = \mathbf{a}^T \mathbf{\Sigma} \mathbf{a}$

Proof: Because $Y - EY$ is univariate, $Y - EY = (Y - EY)^T$. Therefore,

$$\begin{aligned} Var(Y) &= E[(Y - EY)^2] = E[(Y - EY)(Y - EY)^T] \\ &= E[(\mathbf{a}^T \mathbf{X} - \mathbf{a}^T \mu)(\mathbf{a}^T \mathbf{X} - \mathbf{a}^T \mu)^T] = E[\mathbf{a}^T (\mathbf{X} - \mu)(\mathbf{X} - \mu)^T \mathbf{a}] \\ &= \mathbf{a}^T E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]\mathbf{a} = \mathbf{a}^T Cov(\mathbf{X}) a = \mathbf{a}^T \mathbf{\Sigma} a \end{aligned}$$

The last step is due to the definition of covariance matrix.

- The variance of Y depends on the covariance structure of $\mathbf{X}$, as well as the weights given by $\mathbf{a}$.

Linear Combinations of Iris Setosa Features

- Recall that for the iris setosa, $\mathbf{X}$ is 50×4.
- Consider a linear combination of the features $Y = Xb$, where

$$b = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

- Yb is a 50×1 vector, with the ith row be the average of the four features of the ith iris setosa flower. To see this

Linear Combinations of Iris Setosa Features

$$Y = Xb = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} b = \begin{pmatrix} X_1^T b \\ \vdots \\ X_n^T b \end{pmatrix} = \begin{pmatrix} b^T X_1 \\ \vdots \\ b^T X_n \end{pmatrix}$$

$$= \begin{pmatrix} \frac{x_{11}+x_{12}+x_{13}+x_{14}}{4} \\ \vdots \\ \frac{x_{n1}+x_{n2}+x_{n3}+x_{n4}}{4} \end{pmatrix}$$

Linear Combinations of Iris Setosa Features: sample mean

```
setosa=as.matrix(iris[iris$Species=="setosa", 1:4])
sample.meanvec=matrix(colMeans(setosa), 4, 1)
rownames(sample.meanvec)=colnames(setosa)
colnames(sample.meanvec)="mean"
b=matrix(1/4, 4, 1)
Y=setosa%*%b
#sample mean of Y: the following two results are the same
mean(Y)
```

```
## [1] 2.5355
```

```
t(b)%*%sample.meanvec
```

```
##         mean
## [1,] 2.5355
```

Linear Combinations of Iris Setosa Features: sample variance

```
#sample variance of Y: the following two results are the sa
var(Y)
```

```
##             [,1]
## [1,] 0.03844617
```

```
t(b)%*%cov(setosa)%*%b
```

```
##             [,1]
## [1,] 0.03844617
```

Section 3

A Simulated Study

Daily Intake of Protein

- This is a simulated data set
- For adults, the recommended range of daily protein intake is between 0.8 g/kg and 1.8 g/kg of body weight
- 60 observations
- 4 sources of proteins
  - meat
  - dairy
  - vegetables / nuts / tofu
  - other

Choose Mean Vector and Covariance Matrix

- The multivariate distribution has
  - mean vector

$$\boldsymbol{\mu} = \left(24, 16, 8, 8\right)^T$$

  - covariane matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1.3 \end{pmatrix}$$

Define Mean Vector and Covariance Matrix in R

```
#the library "MASS" is required
library(MASS)
my.cov=4*(diag(4) + 0.3* rep(1,4)%o%rep(1,4))
eigen(my.cov)#to check whether the cov matrix is p.d.
```

```
## eigen() decomposition
## $values
## [1] 8.8 4.0 4.0 4.0
##
## $vectors
##      [,1]       [,2]       [,3]       [,4]
## [1,] -0.5  0.8660254  0.0000000  0.0000000
## [2,] -0.5 -0.2886751 -0.5773503 -0.5773503
## [3,] -0.5 -0.2886751 -0.2113249  0.7886751
## [4,] -0.5 -0.2886751  0.7886751 -0.2113249
```

Simulate A Random Sample

```
set.seed(1)
x=mvrnorm(n, mu=my.mean, Sigma=my.cov)
dim(x)
```

```
## [1] 60  4
```

```
protein=as.matrix(data.frame(meat=x[,1],dairy=x[,2],
                             veg=x[,3], other=x[,4]))
```

The simulated data

```
protein
```

```
##          meat    dairy       veg     other
##  [1,] 29.08891 17.54865  5.814221  7.264953
##  [2,] 23.65965 13.06336  8.734581  9.452868
##  [3,] 26.43410 16.83504  9.278807  8.409798
##  [4,] 21.68232 15.51922  3.379171  5.954558
##  [5,] 22.22387 15.45446  8.804571  7.562144
##  [6,] 25.54395 16.46835  8.556332 10.299174
##  [7,] 20.15075 14.71290 10.660378  7.584075
##  [8,] 25.44330 14.98680  4.866275  6.323171
##  [9,] 23.41142 16.34138  6.667006  6.164109
## [10,] 28.21604 16.64242  5.874860  7.078538
## [11,] 22.58127 13.61817  5.178349  5.652878
## [12,] 22.19211 16.04745  8.714666  6.732854
```

Sample Mean and Sample Covariance

```
xbar=matrix(colMeans(protein), 4, 1)
t(xbar)
```

```
##             [,1]     [,2]    [,3]     [,4]
## [1,] 24.03403 15.92836 7.66049 7.738634
```

```
S=cov(protein)
S
```

```
##            meat     dairy       veg     other
## meat  4.2956426 0.8150757 1.1294478 0.5532420
## dairy 0.8150757 4.4052993 0.3497889 0.2337300
## veg   1.1294478 0.3497889 5.1705794 0.5897121
## other 0.5532420 0.2337300 0.5897121 4.5287293
```

Estimation

- An unbiased estimator of $\boldsymbol{\mu}$ is the sample mean vector, i.e., $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$.
- An unbiased estimator of $\boldsymbol{\Sigma}$ is the sample covariance matrix $\mathbf{S}$, i.e., $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$
- We have shown that $Var(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$, where $n = 60$.
- We can estimate it by

$$\hat{Var}(\bar{\mathbf{X}}) = \frac{1}{60}\mathbf{S}$$

Linear Functions/Combinations: Question 1

- Q1: Construct a large-sample (approximate) C.I. for protein from meat. In other words, the parameter of interest is μ_1.
- Estimate $\bar{X}_1 = 24.0$.
- We need compute the standard error (s.e.) of $\bar{X}_1$, wich is defined as $se(\bar{X}_1) = \sqrt{\hat{var}(\bar{X}_1)}$
- Two ways to compute the s.e.,
  1. $se(\bar{X}_1) = \sqrt{4.2956/60} = 0.27$
  2. The calculation can also be done by noticing that $\bar{X}_1$ is a linear combination of $\bar{\mathbf{X}}$: $\bar{X}_1 = \mathbf{a}^T\bar{\mathbf{X}}$, where $\mathbf{a}^T = (1, 0, 0, 0)$. Thus,

$$\hat{Var}(\bar{X}_1) = \mathbf{a}^T \frac{\mathbf{S}}{60} \mathbf{a}$$

Linear Functions/Combinations: Question 1 (continued)

- R code to compute using the above two ways

```
# Method 1
sqrt(S[1,1]/60)
```

```
## [1] 0.2675706
```

```
# Method 2
a=matrix(c(1,0,0,0),4,1)
sqrt(t(a)%*%S%*%a/60)
```

```
##           [,1]
## [1,] 0.2675706
```

- Both methods give $s.e.(\bar{X}_1) = 0.27$
- An approximate 95% C.I. for μ_1 is $24.0 \pm 1.96 * 0.27$

Linear Functions/Combinations: Question 2

- Construct a large-sample C.I. for the total protein intake
- The parameter of interest is $\mu_1 + \mu_2 + \mu_3 + \mu_4 = \mathbf{a}^T \boldsymbol{\mu}$, where $\mathbf{a} = (1, 1, 1, 1)^T$.
- Estimate: $\mathbf{a}^T \bar{\mathbf{X}}$
- Standard error: $\sqrt{\mathbf{a}^T \frac{\mathbf{S}}{n} \mathbf{a}}$

Linear Functions/Combinations: Question 2 (continued)

```
a=matrix(1,4,1)
t(a)%*% xbar #estimate
```

```
##           [,1]
## [1,] 55.36152
```

```
sqrt(t(a)%*%S%*%a/60) #standard error
```

```
##            [,1]
## [1,] 0.6550095
```

```
#a large-sample 95% C.I.
c(t(a)%*% xbar- 1.96*sqrt(t(a)%*%S%*%a/60),
  t(a)%*% xbar+ 1.96*sqrt(t(a)%*%S%*%a/60))
```

```
## [1] 54.07770 56.64534
```

Linear Functions/Combinations: Question 3

- Q3: Construct a large-sample C.I. for the difference of protein intake between from meat and from vegetable
- The parameter of interest is $\mu_1 - \mu_3 = \mathbf{a}^T \boldsymbol{\mu}$, where $\mathbf{a} = (1, -1, 0, 0)^T$.
- Estimate: $\mathbf{a}^T \bar{\mathbf{X}}$
- Standard error: $\sqrt{\mathbf{a}^T \frac{\mathbf{S}}{n} \mathbf{a}}$

Linear Functions/Combinations: Question 3 (continued)

```
a=matrix(c(1,-1,0,0),4,1)
t(a)%*% xbar #estimate
```

```
##          [,1]
## [1,] 8.105671
```

```
sqrt(t(a)%*%S%*%a/60) #standard error
```

```
##           [,1]
## [1,] 0.3432878
```

```
#a large-sample 95% C.I.
c(t(a)%*% xbar- 1.96*sqrt(t(a)%*%S%*%a/60),
  t(a)%*% xbar+ 1.96*sqrt(t(a)%*%S%*%a/60))
```

```
## [1] 7.432827 8.778515
```

Section 4

Generalized Variance

Why Do We Need Generalized Variance?

- For a random variable (i.e., univaraite), we quantity dispersion using variance and standard deviation.
- For a random vector (i.e., multivariate), we use its covariance matrix to quantify the dispersion as well as the relationships between different variables /features.
  - The dispersion information is represented by a matrix, which has $p(p+1)/2$ unique parameters

What is Generalized Variance?

- It is attempting to have a scalar summary (i.e., a single number) to quantify the "total" amount of dispersion for a multivariate distribution
- Generalized variance
  - Provides an overall measure of dispersion of the multivariate distribution
  - One choice is the determinant: $|\mathbf{\Sigma}|$.
  - A larger determinant indicates a greater degree of dispersion

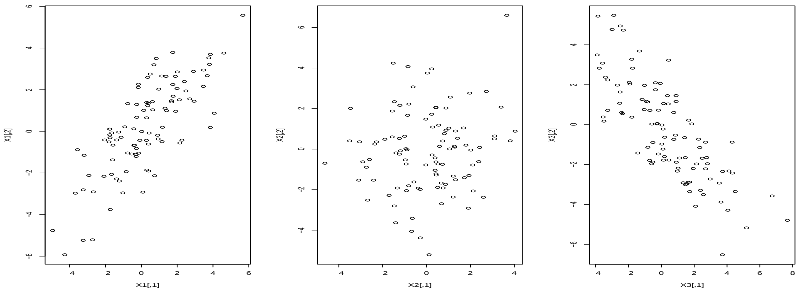
Generalized Inverse: An Example

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}, \mathbf{\Sigma}_3 = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}$$

```
Sigma1=matrix(c(5,4,4,5), 2,2)
X1=mvrnorm(100, mu=c(0,0), Sigma=Sigma1)
Sigma2=matrix(c(3,0,0,3), 2,2)
X2=mvrnorm(100, mu=c(0,0), Sigma=Sigma2)
Sigma3=matrix(c(5,-4,-4,5), 2,2)
X3=mvrnorm(100, mu=c(0,0), Sigma=Sigma3)
```

Generalized Inverse: An Example

```
par(mfrow=c(1,3))
plot(X1); plot(X2); plot(X3)
```



Generalized Variance might (NOT) be useful

- In the example above, $|\Sigma_1| = |\Sigma_2| = |\Sigma_3| = 9$!
- $|\boldsymbol{\Sigma}|$ does not tell the orientations.
- $|\boldsymbol{\Sigma}|$ is useful to compare two patterns when they have nearly the same orientations.
- The generalized variance does not capture all the information contained in the covariance matrix.
- The eigenvalues provide more information than the determinant - Principal Component Analysis!

How to Interpret A Covariance Matrix? - the 2D Situation

https://datascienceplus.com/understanding-the-covariance-matrix/

Section 5

Normal Distributions: univariate, multivariate, matrix normal distributions

The Big Picture: Univariate vs Multivariate

- Review: A random sample, denoted by $X_1, \cdots, X_n$, from a (univariate) normal distribution $N(\mu, \sigma^2)$
  - What are the distributions of $\bar{X}, s^2$? What useful statistics can be constructed?
- New material: A random sample, denoted by $\mathbf{X}_1, \cdots, \mathbf{X}_n$, from a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - What are the distributions of $\bar{\mathbf{X}}, \mathbf{S}^2$? What useful statistics can be constructed?

The Big Picture: Univariate

- A random sample, denoted by $X_1, \cdots, X_n$, from a (univariate) normal distribution $N(\mu, \sigma^2)$
- Let $\mathbf{X}_{n\times 1} = (X_1, \cdots, X_n)^T$. It is random vector with a multivarite normal distribution, i.e.,

$$\mathbf{X}_{n\times 1} = (X_1, \cdots, X_n)^T \sim \mathbf{N}(\mu\mathbf{1}, \sigma^2\mathbf{I})$$

1. $\bar{X} \sim N(\mu, \sigma^2/n)$
2. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
3. a t-statistic is

$$\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{s}$$

It follows the t-distribution with n-1 degrees of freedom, denoted by t_{n-1}.

The Big Picture: Multivariate

- A random sample $\mathbf{X}_1, \cdots, \mathbf{X}_n$ from a multivariate normal distribution $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Let

$$\mathbf{X}_{n \times p} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix}$$

$\mathbf{X}$ follows a matrix normal distribution.

1. Sample mean vector follows a multivariate normal, i.e., $\bar{\mathbf{X}} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$
2. Sample covariance matrix $(n-1)\mathbf{S}$ follows a Wishart distribution, i.e., $(n-1)\mathbf{S} \sim Wishart_p(n-1, \Sigma)$
3. Hoetelling's T^2

$$T^2 = (\bar{\mathbf{X}} - \boldsymbol{\mu})^T \left(\frac{\mathbf{S}}{n}\right)^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$$