

# Lecture 3

## Multivariate Normal Distribution, Copulas, and Nonparametric Density Estimation

Readings: Zelterman, 2015 Chapters 5, 6, 7, Izeman, 2008 Chapter 4.1, 4.3, 4.5

DSA 8070 Multivariate Analysis

Whitney Huang  
Clemson University



### Notes

---

---

---

---

---

---

### Agenda

1 Multivariate Normal Distribution

2 Geometry of the Multivariate Normal Density

3 Copula Models

4 Nonparametric Density Estimation



### Notes

---

---

---

---

---

---

### The Multivariate Normal Distribution

As the **univariate normal** is central to **univariate statistics**, the **multivariate normal** is central to **multivariate statistics**

- **Mathematical Simplicity:** It is easy to obtain multivariate methods based on the multivariate normal distribution
- **Central Limit Theorem:** The *sample mean vector is going to be approximately multivariate normally distributed when the sample size is sufficiently large*
- Many natural phenomena may be modeled using this distribution (perhaps after transformation)



### Notes

---

---

---

---

---

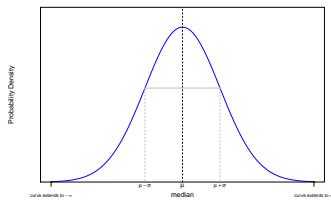
---

## Review: Univariate Normal Distributions

The probability density function of the normal distribution is

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

where  $\mu$  and  $\sigma^2$  are its **mean** and **variance**, respectively.



$\left(\frac{x-\mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$  is the squared statistical distance between  $x$  and  $\mu$  in standard deviation units

## Notes

---



---



---



---



---



---

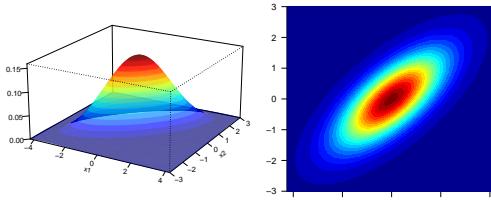


---

## Multivariate Normal Distributions

If we have a  $p$ -dimensional random vector that is distributed according to a **multivariate normal distribution** with mean vector  $\mu = (\mu_1, \mu_2, \dots, \mu_p)^T$  and covariance matrix  $\Sigma = \{\sigma_{ij}\}$ , the probability density function is

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



## Notes

---



---



---



---



---



---



---

## Review: Central Limit Theorem (CLT)

The **sampling distribution** of the **mean** will become approximately **normally distributed** as the **sample size becomes larger, irrespective of the shape of the population distribution!**

Let  $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} F$  with  $\mu = E[X_i]$  and  $\sigma^2 = \text{Var}[X_i]$ . Then  $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$  as  $n \rightarrow \infty$ .

## Notes

---



---



---



---



---



---



---

## CLT In Action

- ① Generate 100 ( $n$ ) random numbers from an Exponential distribution (population distribution)
  - ② Compute the **sample mean** of these 100 random numbers
  - ③ Repeat this process 120 times



## Notes

---

---

---

---

---

---

---

---

---

---

## Properties of the Multivariate Normal Distribution

- If  $X \sim N(\mu, \Sigma)$ , then any subset of  $X$  also has a multivariate normal distribution

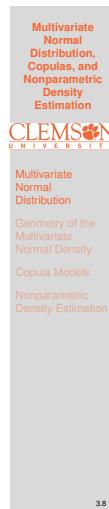
**Example:** Each single variable  $X_i \sim N(\mu_i, \sigma_i^2)$ ,  $i = 1, \dots, p$

- If  $X \sim N(\mu, \Sigma)$ , then any linear combination of the variables has a univariate normal distribution

**Example:** If  $Y = \mathbf{a}^T \mathbf{X}$ . Then  $Y \sim N(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})$

- Any conditional distribution for a subset of the variables conditional on known values for another subset of variables is a multivariate distribution

**Example:**  $X_1|X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$



## Notes

---

---

---

---

---

---

---

**Example: Linear Combination of the Cholesterol Measurements [source: Penn State Univ. STAT 505]**

Cholesterol levels were taken 0, 2, and 4 days following the heart attack on  $n$  patients. The mean vector is:

	Variable	Mean
$\bar{x}$	$X_1$ (0-day)	259.5
	$X_2$ (2-day)	230.8
	$X_3$ (4-day)	221.5

and the covariance matrix

$$S = \begin{bmatrix} 2276 & 1508 & 813 \\ 1508 & 2206 & 1349 \\ 813 & 1349 & 1865 \end{bmatrix}$$

Suppose we are interested in  $\Delta = X_2 - X_1$ , the difference between the 2-day and the 0-day measurements. We can write the linear combination of interest as

$$\Delta = \mathbf{a}^T \mathbf{X} = [-1 \quad 1 \quad 0] \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$



## Notes

---

---

---

---

---

---

---

## Cholesterol Measurements Example Cont'd

- The mean value for the difference  $\Delta$  is

$$[-1 \ 1 \ 0] \begin{bmatrix} 259.5 \\ 230.8 \\ 221.5 \end{bmatrix} = -28.7$$

- The variance for  $\Delta$  is

$$\begin{aligned} & [-1 \ 1 \ 0] \begin{bmatrix} 2276 & 1508 & 813 \\ 1508 & 2206 & 1349 \\ 813 & 1349 & 1865 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \\ &= [-768 \ 698 \ 536] \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \\ &= 1466 \end{aligned}$$

- If we assume these three variables together follows a multivariate normal distribution, then  $\Delta$  follows a univariate normal distribution



## Notes

---



---



---



---



---



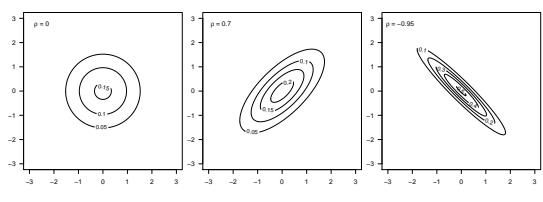
---

## Bivariate Normal Distribution

Let's focus bivariate normal distributions first as we can visualize them to facilitate our understanding. Suppose we have  $X_1$  and  $X_2$  jointly follows a bivariate normal distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right]$$

Let's fix  $\mu_1 = \mu_2 = 0$  and  $\sigma_1^2 = \sigma_2^2 = 1$



## Notes

---



---



---



---



---



---

## Exponent of Multivariate Normal Distribution

Recall the multivariate normal density:

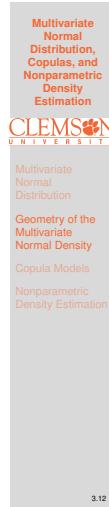
$$f(x) = \frac{1}{2\pi^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right\}.$$

This density function only depends on  $x$  through the squared Mahalanobis distance:  $(x - \mu)^T \Sigma^{-1} (x - \mu)$

- For bivariate normal, we get an ellipse whose equation is  $(x - \mu)^T \Sigma^{-1} (x - \mu) = c^2$  which gives all  $x = (x_1, x_2)$  pairs with constant density

- These ellipses are call contours and all are centered around  $\mu$

- A constant probability contour equals
  - $=$  all  $x$  such that  $(x - \mu)^T \Sigma^{-1} (x - \mu) = c^2$
  - $=$  surface of ellipsoid centered at  $\mu$



## Notes

---



---



---



---



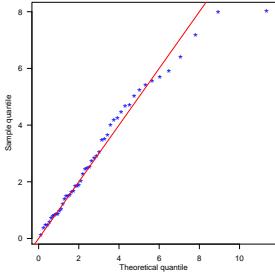
---



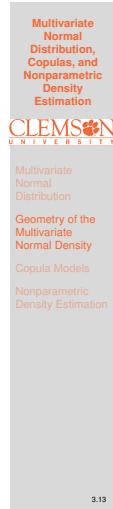
---

## Multivariate Normality and Outliers

The variable  $d^2 = (\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})$  has a chi-square distribution with  $p$  degrees of freedom, i.e.,  $d^2 \sim \chi_p^2$  if  $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \Rightarrow$  we can exploit this result to check multivariate normality and to detect outliers



- Sort  $(\mathbf{x}_i - \boldsymbol{\bar{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \boldsymbol{\bar{x}})$  in an increasing order to get sample quantiles
- Calculate the theoretical quantiles using the chi-square quantiles with  $p = \frac{i-0.5}{n}$ ,  $i = 1, \dots, n$
- Plot sample quantile against theoretical quantiles



## Notes

---

---

---

---

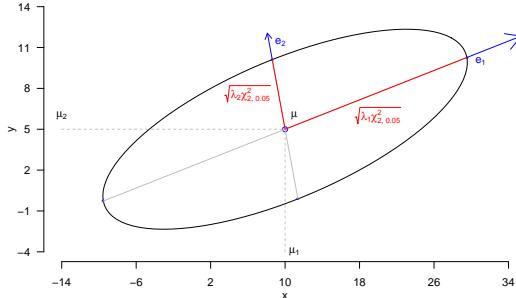
---

---

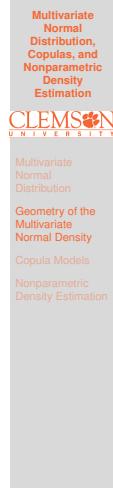
## Eigen-Structure of $\Sigma$ and the Geometry of the Multivariate Normal Density

Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ , where  $\boldsymbol{\mu} = (10, 5)^T$  and  $\Sigma = \begin{bmatrix} 64 & 16 \\ 16 & 9 \end{bmatrix}$ .

The 95% probability contour is shown below



Next, we talk about how to "draw" this contour



## Notes

---

---

---

---

---

---

## Probability Contours

- The solid ellipsoid of values  $x$  satisfy

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq c^2 = \chi_{df=p,\alpha}^2$$

Here we have  $p = 2$  and

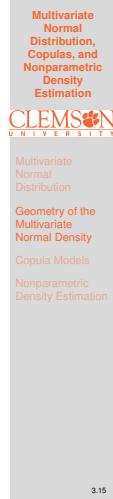
$$\alpha = 0.05 \Rightarrow c = \sqrt{\chi_{2,0.05}^2} = 2.4478$$

- Major axis:  $\boldsymbol{\mu} \pm c\sqrt{\lambda_1} \mathbf{e}_1$ , where  $(\lambda_1, \mathbf{e}_1)$  is the first eigenvalue/eigenvector of  $\Sigma$ .

$$\Rightarrow \lambda_1 = 68.316, \quad \mathbf{e}_1 = \begin{bmatrix} -0.9655 \\ -0.2604 \end{bmatrix}$$

- Minor axis:  $\boldsymbol{\mu} \pm c\sqrt{\lambda_2} \mathbf{e}_2$ , where  $(\lambda_2, \mathbf{e}_2)$  is the second eigenvalue/eigenvector of  $\Sigma$ .

$$\Rightarrow \lambda_2 = 4.684, \quad \mathbf{e}_2 = \begin{bmatrix} 0.2604 \\ -0.9655 \end{bmatrix}$$



## Notes

---

---

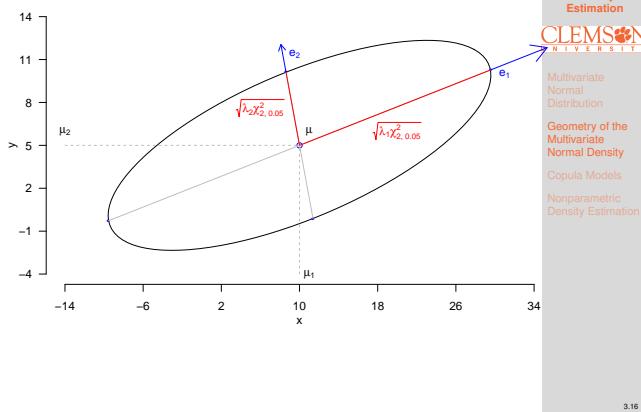
---

---

---

---

## Graph of 95% Probability Contour



## Notes

---

---

---

---

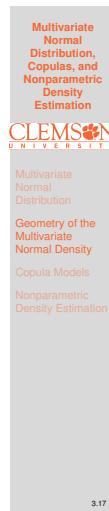
---

---

### Example: Wechsler Adult Intelligence Scale [source: Penn State Univ. STAT 505]

We have data (wechslet.txt) on 37 subjects ( $n = 37$ ) taking the Wechsler Adult Intelligence Test, which consists four different components: 1) Information; 2) Similarities; 3) Arithmetic; 4) Picture Completion.

- ① Calculate the sample mean vector  $\bar{x}$  and covariance matrix  $S$
- ② Compute the eigenvalues and eigenvectors of  $S$  and give a geometry interpretation
- ③ Diagnostic the multivariate normal assumption



## Notes

---

---

---

---

---

---

### Beyond Normality: Copulas [Sklar, 1959; Joe, 1997]

#### Sklar's theorem:

$$\begin{aligned} F(x_1, \dots, x_p) &= \mathbb{P}r(X_1 \leq x_1, \dots, X_p \leq x_p) \\ &= \mathbb{P}r(F_1^{-1}(U_1) \leq x_1, \dots, F_p^{-1}(U_p) \leq x_p) \\ &= \mathbb{P}r(U_1 \leq F_1(x_1), \dots, U_p \leq F_p(x_p)) \\ &= C(\underbrace{F_1(x_1), \dots, F_p(x_p)}_{\text{copula}}) \end{aligned}$$

⇒ a copula is a multivariate CDF with Uniform(0, 1) margins

#### Multivariate Normal

#### Copulas

- |   |   |
|---|---|
| • Margins: Normal                                       | • Margins: arbitrary $F_j$                            |
| • Dependence: $\Sigma$ (elliptical; correlation-driven) | • Dependence: copula $C$ (tail-dependent, asymmetric) |
| • Tails: light; no tail dependence                      | • Tails: flexible                                     |

Copula approach has becomes popular in many areas, e.g., quantitative finance as it allows for **separate modeling of marginal distributions and dependence structure**



## Notes

---

---

---

---

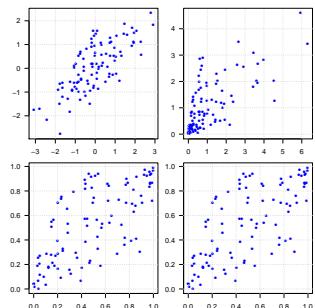
---

---

## An Illustration of Bivariate Gaussian Copula

**Left:** Normal marginals + Gaussian Copula ( $\rho = 0.7$ )

**Right:** Exponential marginals + Gaussian Copula ( $\rho = 0.7$ )



The copula approach allows us to “build” multivariate distributions with non-normal marginals

Multivariate  
Normal  
Distribution,  
Copulas, and  
Nonparametric  
Density  
Estimation

CLEMSON  
UNIVERSITY

Multivariate  
Normal  
Distribution

Geometry of the  
Multivariate  
Normal Density

Copula Models

Nonparametric  
Density Estimation

3.19

## Notes

---

---

---

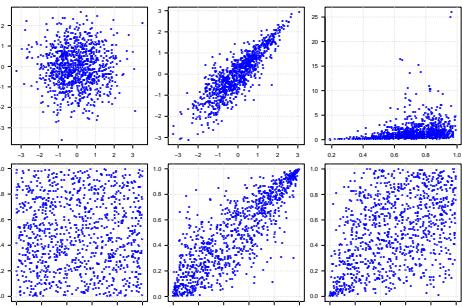
---

---

---

## More Examples

**Marginal:** normal and normal  
**Copula:** Gaussian  
 $\rho = 0$



**Marginal:** normal and normal  
**Copula:** Gumbel  
 $\theta = 3$

Multivariate  
Normal  
Distribution,  
Copulas, and  
Nonparametric  
Density  
Estimation

CLEMSON  
UNIVERSITY

Multivariate  
Normal  
Distribution

Geometry of the  
Multivariate  
Normal Density

Copula Models

Nonparametric  
Density Estimation

3.20

## Notes

---

---

---

---

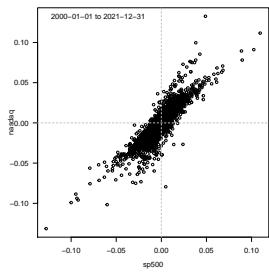
---

---

⇒ The copula approach allows for more options for dependence modeling

## A Financial Application Using Copula

Here we illustrate how to use a copula to model the bivariate joint distribution of S&P 500 and Nasdaq (log) returns



- ① Transform the data  $(x_{1i}, x_{2i})_{i=1}^n$  to  $(u_{1i}, u_{2i})_{i=1}^n$  and fit a copula model to it
- ② Fit a distribution to  $\{x_{1i}\}_{i=1}^n$  and  $\{x_{2i}\}_{i=1}^n$ , respectively
- ③ Combine the fitted copula and marginal distributions to form the fitted bivariate distribution

Multivariate  
Normal  
Distribution,  
Copulas, and  
Nonparametric  
Density  
Estimation

CLEMSON  
UNIVERSITY

Multivariate  
Normal  
Distribution

Geometry of the  
Multivariate  
Normal Density

Copula Models

Nonparametric  
Density Estimation

3.21

## Notes

---

---

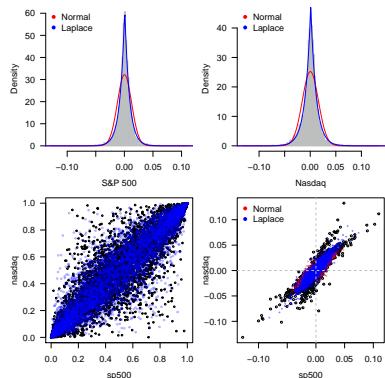
---

---

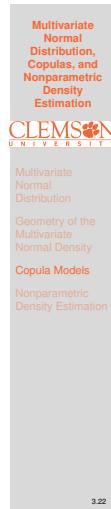
---

---

## Marginals, Copula, and Joint Distribution



**Marginals:** Laplace    **Copula:**  $t$ -copula



## Notes

---

---

---

---

---

---

---

---

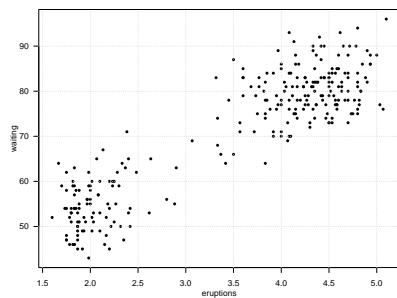
---

---

---

## Old Faithful Geyser Data

Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone NP



Sometimes, parametric models may not be sufficient to model multivariate data



## Notes

---

---

---

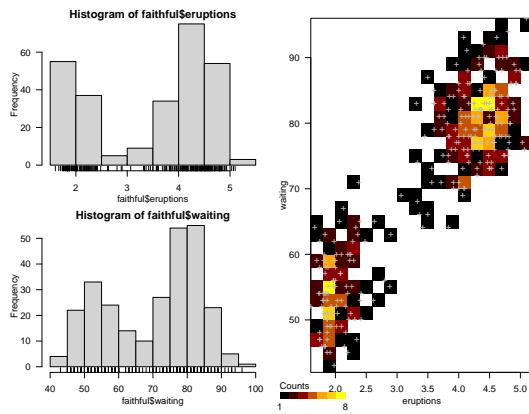
---

---

---

---

## Histograms of Old Faithful Data



## Notes

---

---

---

---

---

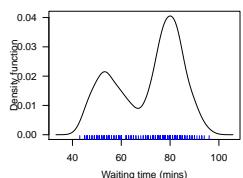
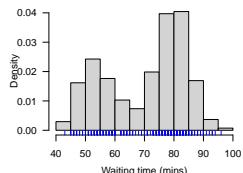
---

---

## Transition from Histogram to Kernel Density

**Goal:** to estimate the probability density function  $f(x)$

- **Histogram:**



- **Kernel Density:**

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right),$$

where  $K(\cdot)$  is the kernel function and  $h$  is the bandwidth

Multivariate Normal Distribution, Copulas, and Nonparametric Density Estimation  
CLEMSON UNIVERSITY  
Multivariate Normal Distribution  
Geometry of the Multivariate Normal Density  
Copula Models  
Nonparametric Density Estimation  
3.25

## Notes

---

---

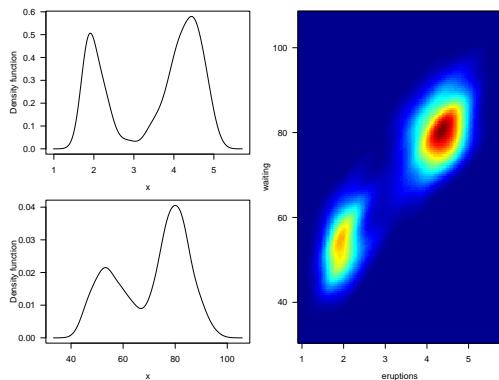
---

---

---

---

## Kernel Density Estimates of Old Faithful



Multivariate Normal Distribution, Copulas, and Nonparametric Density Estimation  
CLEMSON UNIVERSITY  
Multivariate Normal Distribution  
Geometry of the Multivariate Normal Density  
Copula Models  
Nonparametric Density Estimation  
3.26

## Notes

---

---

---

---

---

---

The choice of bandwidth plays a crucial role (try it yourself in an R session)

## Summary

In this lecture, we learned about:

- Multivariate Normal Distribution
- Copula Modeling
- Non-parametric Density Estimation

In the next lecture, we will study inference for a mean vector and for multiple mean vectors

Multivariate Normal Distribution, Copulas, and Nonparametric Density Estimation  
CLEMSON UNIVERSITY  
Multivariate Normal Distribution  
Geometry of the Multivariate Normal Density  
Copula Models  
Nonparametric Density Estimation  
3.27

## Notes

---

---

---

---

---

---