Lecture 4

Multivariate Normal Distribution, Copula, and Nonparametric Density Estimation

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

DSA 8070 Multivariate Analysis
September 12 - September 16, 2022

Whitney Huang
Clemson University

Agenda

1. Multivariate Normal Distribution
2. Geometry of the Multivariate Normal Density
3. Copula
4. Nonparametric Density Estimation

The Multivariate Normal Distribution

Just as the univariate normal distribution tends to be the most important distribution in univariate statistics, the multivariate normal distribution is the most important distribution in multivariate statistics

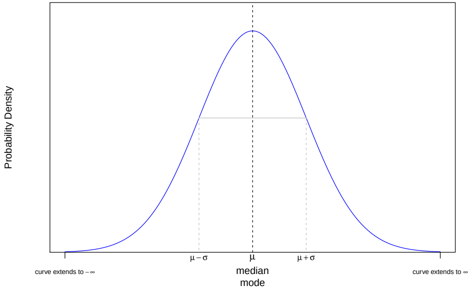
- **Mathematical Simplicity**: It is easy to obtain multivariate methods based on the multivariate normal distribution
- **Central Limit Theorem**: *The sample mean vector is going to be approximately multivariate normally distributed when the sample size is sufficiently large*
- Many natural phenomena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

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 μ and σ^2 are its mean and variance, respectively.

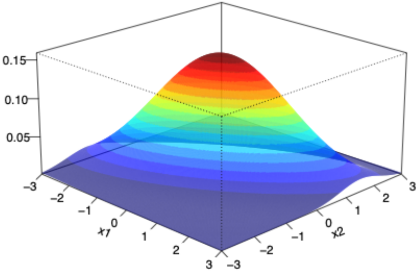


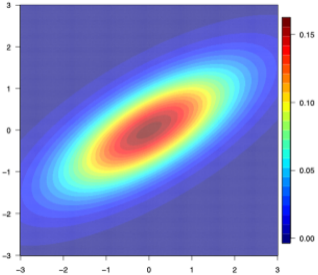
$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$ is the squared statistical distance between x and μ in standard deviation units

Multivariate Normal Distributions

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

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





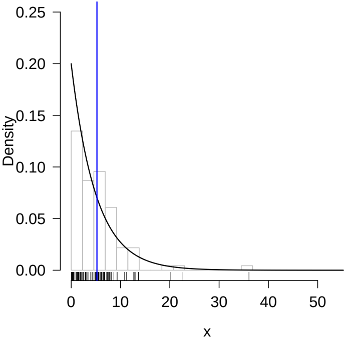
Review: Central Limit Theorem (CLT)

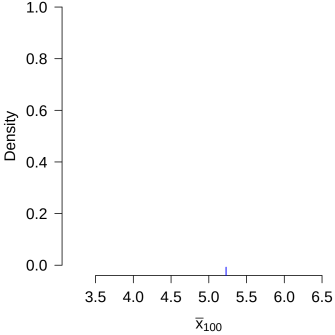
The **sampling distribution** of the **mean** will become approximately **normally distributed** as the **sample size becomes larger**, **irrespective of the shape of the population distribution!**

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

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





Properties of the Multivariate Normal Distribition

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

  Example: Each single variable $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2), \quad i = 1, \cdots, p$

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

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

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

  Example:
  $\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$

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:

$\bar{\boldsymbol{x}}$ =

Variable	Mean
X_1 (0-day)	259.5
X_2 (2-day)	230.8
X_3 (4-day)	221.5

and the covariance matrix

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

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

$$\Delta = \boldsymbol{a}^T \boldsymbol{X} = \begin{bmatrix} -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

Cholesterol Measurements Example Cont'd

- The mean value for the difference Δ is

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

- The variance for Δ is

$$\begin{bmatrix} -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2276 & 1508 & 813 \\ 1508 & 2206 & 1349 \\ 813 & 1349 & 1865 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} -768 & 698 & 536 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

$$= 1466$$

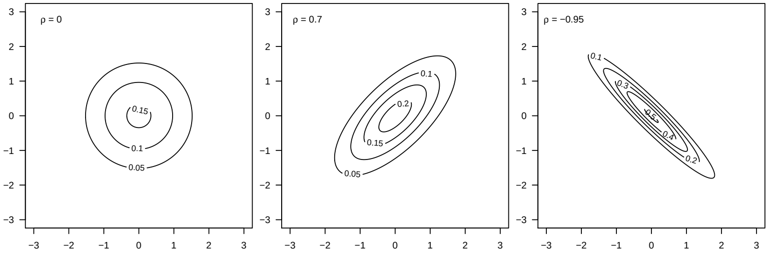
- If we assume these three variables together follows a multivariate normal distribution, then Δ follows a univariate normal distribution

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:

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

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



Exponent of Multivariate Normal Distribution

Recall the multivariate normal density:

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

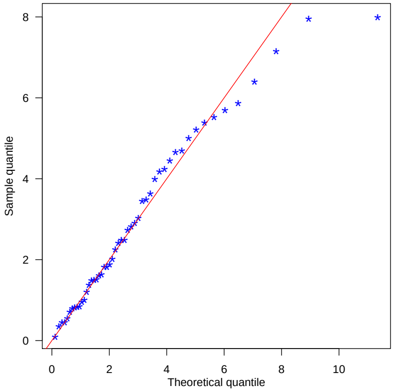
This density function only depends on $\boldsymbol{x}$ through the squared Mahalanobis distance: $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$

- For bivariate normal, we get an ellipse whose equation is $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2$ which gives all $\boldsymbol{x} = (x_1, x_2)$ pairs with constant density
- These ellipses are call contours and all are centered around $\boldsymbol{\mu}$
- A constant probability contour equals

$$\begin{aligned} &= \text{all } \boldsymbol{x} \text{ such that } (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c^2 \\ &= \text{surface of ellipsoid centered at } \boldsymbol{\mu} \end{aligned}$$

Multivariate Normality and Outliers

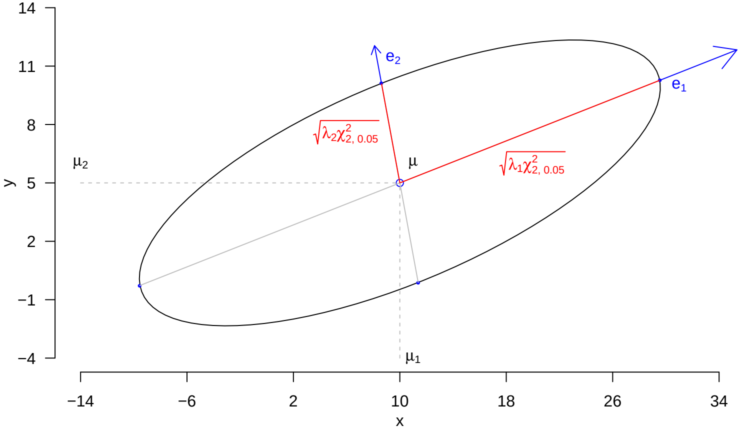
The variable $d^2 = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom , i.e., $d^2 \sim \chi^2_p$ if $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow$ we can exploit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



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

Eigenvalues and Eigenvectors of Σ and the Geometry of the Multivariate Normal Density

Let $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (10, 5)^T$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 64 & 16 \\ 16 & 9 \end{bmatrix}$. The 95% probability contour is shown below



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

Probability Contours

- The solid ellipsoid of values x satisfy

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \le c^2 = \chi^2_{df=p,\alpha}$$

Here we have $p = 2$ and $\alpha = 0.05 \Rightarrow c = \sqrt{\chi^2_{2,0.05}} = 2.4478$

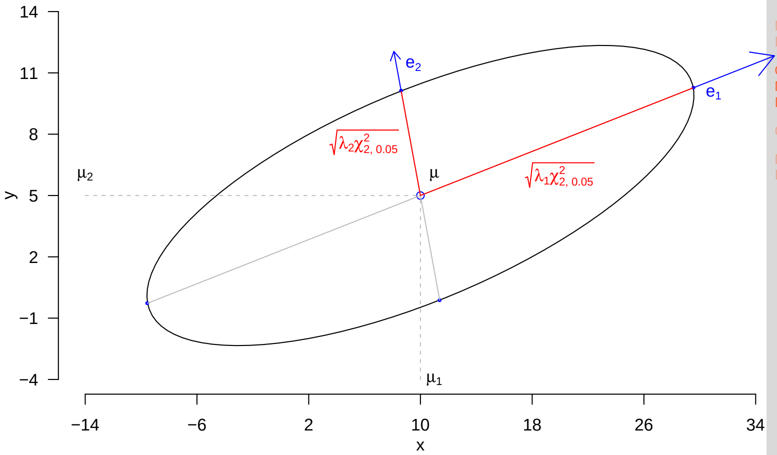
- Major axis: $\boldsymbol{\mu} \pm c\sqrt{\lambda_1 \boldsymbol{e}_1}$, where $(\lambda_1, \boldsymbol{e}_1)$ is the first eigenvalue/eigenvector of $\boldsymbol{\Sigma}$.

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

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

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

Graph of 95% Probability Contour



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 $\bar{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{S}$
2. Compute the eigenvalues and eigenvectors of $\boldsymbol{S}$ and give a geometry interpretation
3. Diagnostic the multivariate normal assumption

Beyond Normality: Copula [Sklar, 1959; Joe, 1997]

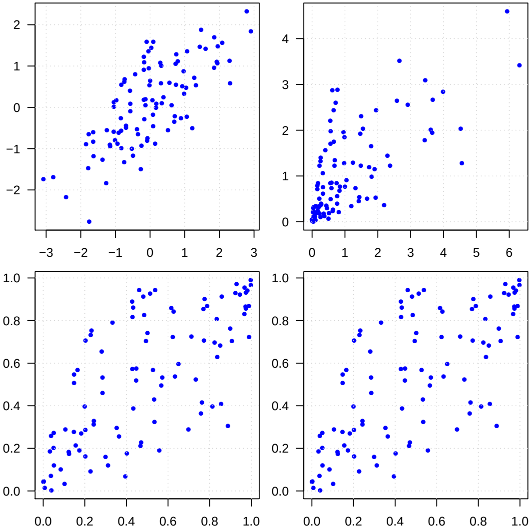
A copula is a multivariate cumulative distribution function for which the marginal probability distribution of each variable is uniform on the interval $[0, 1]$

$$\begin{aligned} F(x_1, \cdots, x_p) &= \mathbb{Pr}(X_1 \le x_1, \cdots, X_p \le x_p) \\ &= \mathbb{Pr}(F_1^{-1}(U_1) \le x_1, \cdots, F_p^{-1}(U_p) \le x_p) \\ &= \mathbb{Pr}(U_1 \le F_1(x_1), \cdots, U_p \le F_p(x_p)) \\ &= C\left(F_1(x_1), \cdots, F_p(x_p)\right) \end{aligned}$$

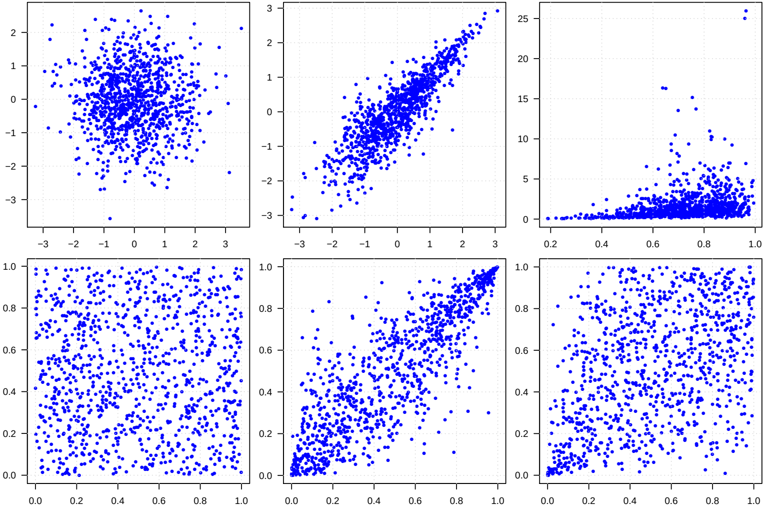
- Copulas are used to model the dependence between random variables
- Copula approach has becomes popular in many areas, e.g., quantitative finance as it allows for separate modeling of marginal distributions and dependence structure

An Illustration of Bivariate Gaussian Copula

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

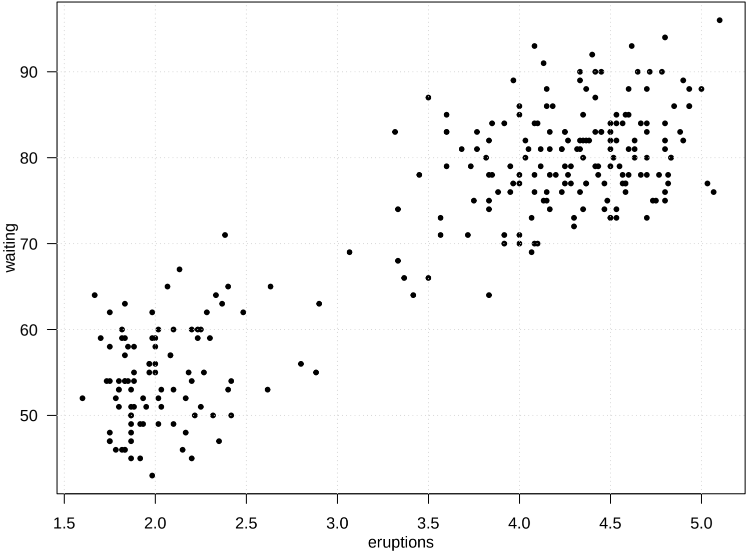


More Examples

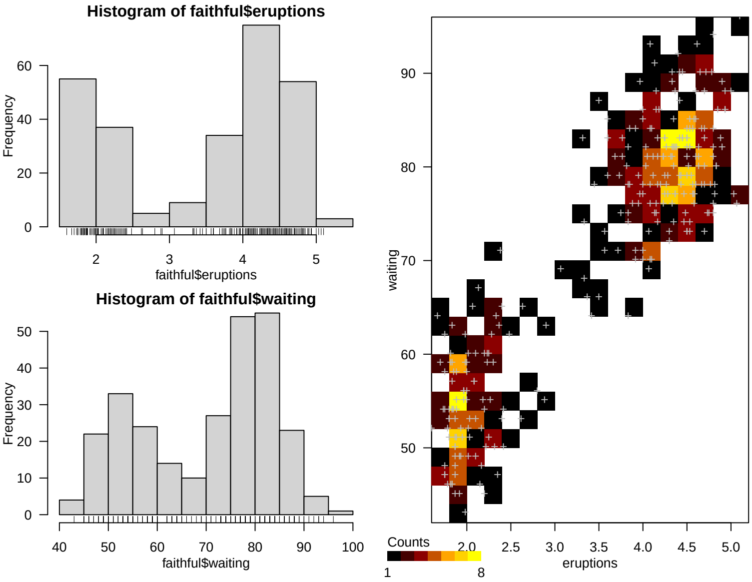


Old Faithful Geyser Data

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



Hisograms of Old Faithful Data



Kernel Density Estimates of Old Faithful