Display Multivariate Data and Measures of Distance

Instructor: Wanhua Su

STAT 372, Covers Chapters 3 and 5 from the Textbook

This note covers four main topics.

- How to display multivariate data? We will introduce scatterplot, scatterplot matrix, growth curve, stars plot, and Chernoff faces.
- How to measure distance between multivariate observations? Manhattan distance, Euclidean distance, Mahalanobis distance, Hamming distance will be covered.
- Multivariate normal distribution. We focus on bivariate normal in this course.
- Sampling distribution of the sample mean vector $\bar{\mathbf{X}}$ and the sample covariance matrix $\boldsymbol{S}$.

Learning Outcomes

After finishing this chapter, students should be able to

- Plot the index scatter plots matrix, draw a stars plot and Chernoff faces plot of a given multivariate data set using R, and interpret the plots.
- Verify whether a distance function is valid.
- Choose the proper distance metric to calculate the distance between observations based on the data types for multivariate data.
- Write down the density function of a multivariate normal distribution.
- Describe the properties of a multivariate normal distribution.
- Find a $(1-\alpha) \times 100\%$ contour for a given bivariate normal distribution.
- Explain the distributions related to the sample mean vector and sample covariance matrix.

1 Display Multivariate Data

In multivariate analysis, at least two measurements are taken from the same individuals. Before conducting any data analysis, we should examine the preliminary relationship among the data using graphs. We will cover several popular ways to display multivariate data.

1.1 Scatterplot

We have seen the 2-demensional scatter plot in Stat 151 when two measurements are taken on the same individuals, for example, age and price of a used car. In a scatter plot, we put one variable in the x-axis and another one in the y-axis. From a scatter plot, we can tell the direction, form and strength of the relationship between the two variables.

If three measurements are taken from each individual, 3-demensional scatter plot can be used with one variable on the x, y, z axis respectively. Take the Iris flowers data for example, a 2-dimensional scatter plot based on "sepal width" and "petal width" is given on the left panel of Figure 1 and a 3-dimensional scatter plot bases on "sepal width", "petal width" and "sepal length" is given on the right panel of Figure (1). Different symbols are used for three different species: red circle for Setosa, black plus for Versicolor, and green cross for Virginica. A plot with different symbols for distinct groups is called an *index* plot.

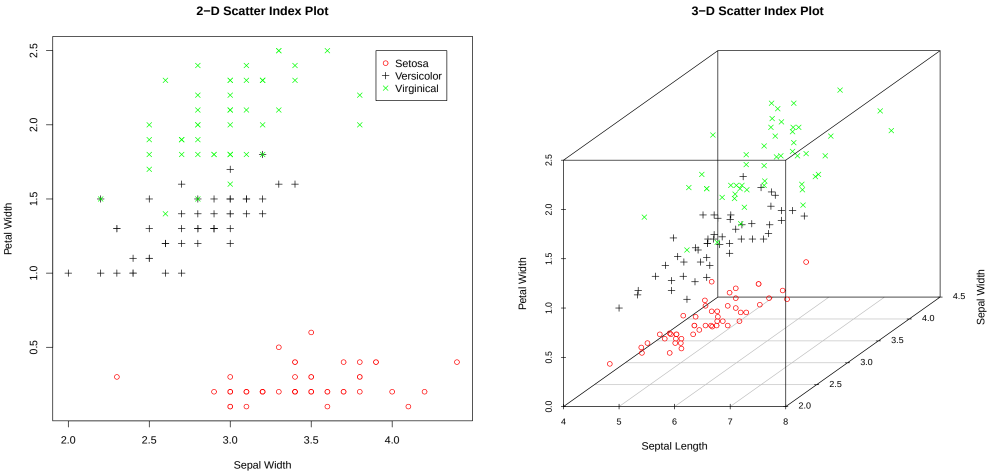

Figure 1: Left: 2-D scatter index plot. Right: 3-D scatter index plot. Three species—red circle: Setosa, black "+": Versicolor, green "x": Virginica

If there are more than 3 measurements, a matrix of scatter plots is used to explore the relationship between any two variables. Figure (2) shows the scatter plots matrix for the Iris flowers data.

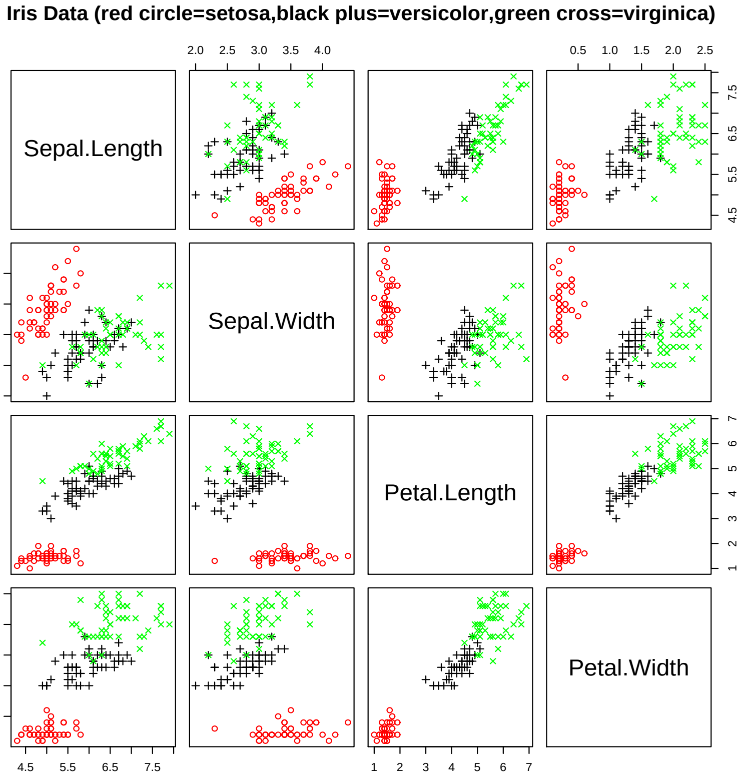

Figure 2: Scatter plots matrix for the Iris flowers data

The scatter plots suggest that the variable "petal width" might be a useful variable which separates the three species.

1.2 Graphs of Growth Curves

When the height of a child is measured at each birthday, the points can be plotted and connected by lines to produce a graph. This is an example of *growth curve* which is widely used for repeated measurements of the same characteristic on the same individuals at different visits. Figure 3 shows the reading ability of six kids at two different ages. Each individual measured twice; therefore, the measurements are repeated measurements. If this information is ignored, a scatter plot of "reading ability" versus "age" (left panel) suggests a negative association; that means reading ability drops when the kids get older, this is intuitively not true. The growth curve connects the two measurements on the same kids and it shows a positive association, that is reading ability grows when the kids get older.

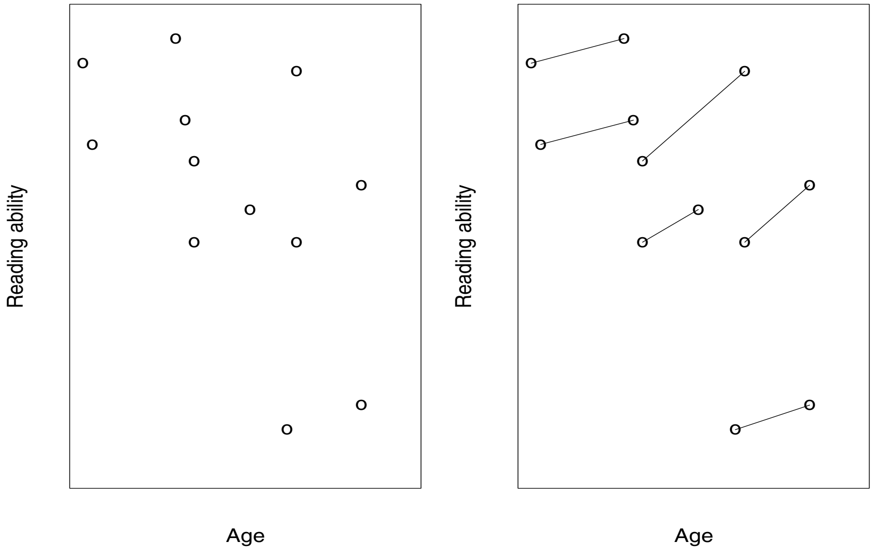

Figure 3: Growth curves of kid's reading ability at two different ages

1.3 Star Plots

A *star plot* (or called radar chart) is a plot that consists of a sequence of equi-angular rays, with each ray representing one of the variables. The data length of a ray is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each ray. This gives the plot a star-like appearance and the origin of one of the popular names for this plot. The star plot can be used to answer the following questions:

- Which observations are most similar? Are there clusters of observations?
- Are there outliers?
- What is the trend of change along time for repeated measurements?

Figure 4 is the stars plot of 36 randomly picked iris flowers, 12 from each species. The numbers under the stars plot are the ID number of the flower, we know that observations 1 to 50 belong to Setosa, 51 to 100 belong to Versicolor and 101 to 150 belong to Virginica. Could you tell how many clusters of flowers?

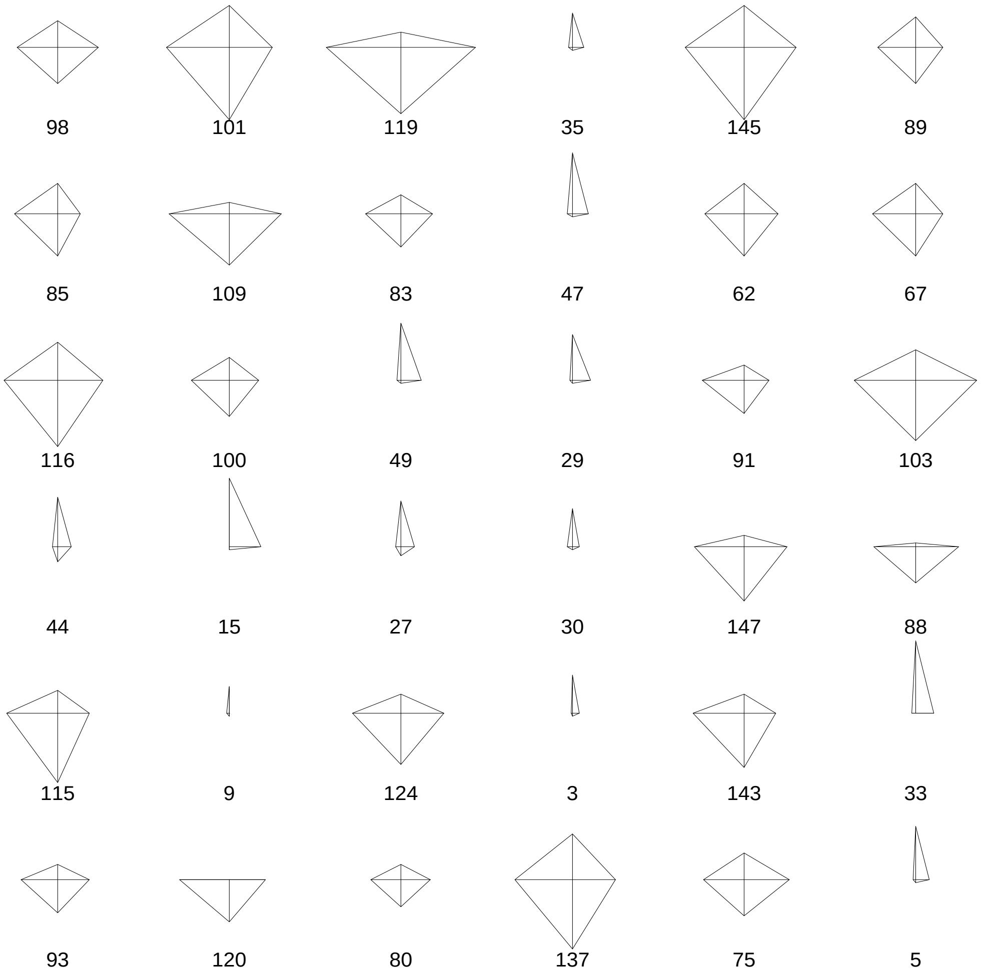

Figure 4: Stars plot of 36 observations of the Iris flowers data

1.4 Chernoff Faces Plot

Chernoff faces, proposed by Herman Chernoff, is a novel method of representing multivariate data by a cartoon of a face whose features, such as length of nose and curvature of mouth, correspond to the variables. Thus each multivariate observation is visualized as a computer-drawn face. The implementation of Chernoff faces plot in R is able to take up to 15 variables, those variables correspond to the features of the face such as: height of face, width of face, structure of face, height of mouth, width of mouth, smiling, height of eyes, width of eyes, height of hair, width of hair, style of hair, height of nose, width of nose, width of ear, height of ear.

Similar to stars plot, Chernoff faces are useful for identifying different groups (clusters) and showing changes over time for repeated measurements. Figure 5 shows the faces of 36 randomly picked iris flowers, 12 from each species. The 15 features of the faces correspond to sepal length, sepal width, petal length, petal width in cycles.

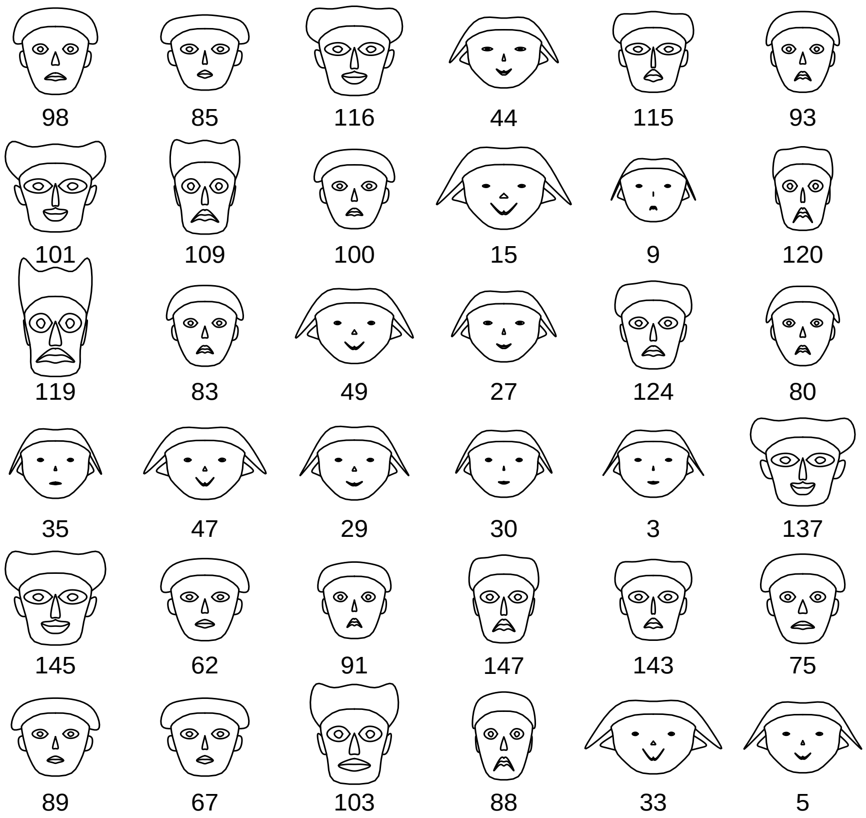

Figure 5: Chernoff faces of 36 observations of the Iris flowers data

2 Distance in Multivariate Analysis

In multivariate analysis, most methods are based on the simple concept of distance. Take clustering analysis for example, we need to group observations that are similar or close to one another. Therefore, we need to calculate the distance between the observations.

2.1 Distances for Quantitative Variables

In univariate cases, the distance between two observations x_1 and x_2 is defined as $d(x_1, x_2) = |x_1 - x_2| = \sqrt{(x_1 - x_2)^2}$. This definition can be extended to the multivariate cases. Suppose $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ and $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$ are two vectors, their *Euclidean* distance is defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{(\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})}.$$

And the *Manhattan* distance is defined as

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| = \sum_{i=1}^{n} |x_i - y_i|.$$

The *Minkowski* distance is defined as

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p},$$

which includes the Manhattan distance (when $p = 1$) and the Euclidean distance (when $p = 2$) as special cases.

Euclidean distance treats each coordinate equally without accounting for the amount of variability in each dimension. A measure that does take into account the variance and covariance of the variables is the *Mahalanobis* distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix which can be replaced by the sample variance-covariance matrix if it is unknown.

One can define his own way to calculate distance as long as the function $d(.)$ satisfies the following properties:

1. Non-negative. For any $\mathbf{x}$ and $\mathbf{y}$, $d(\mathbf{x}, \mathbf{y}) \geq 0$.
2. Identified. $d(\mathbf{x}, \mathbf{x}) = 0$.
3. Symmetric. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$.
4. Definite. If $d(\mathbf{x}, \mathbf{y}) = 0$, then $\mathbf{x} = \mathbf{y}$.
5. Triangle inequality. For any $\mathbf{x} \neq \mathbf{y}, \mathbf{z}$, $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

Example: Verify that the Euclidean distance is a valid distance merit.

Note: If a "unit change" means dramatically different things for different variables, we shall standardize the measurements by subtracting its mean and dividing its standard deviation before we calculate the distance.

2.2 Distance for Categorical Variables

For categorical variables whose values are categories, it is meaningless to calculate the distance using the functions given in the previous section. For example, there are four possible blood types: A, B, O, AB. Even though we can recode the values as 1=A, 2=B, 3=O and 4=AB, we would not say that the distance between types A and B is closer than the distance between types A and O. For categorical measurements, we can use the *Hamming* distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \mathrm{I}(x_i \neq y_i),$$

where I(.) is an indicator function which takes value 1 if the statement is true otherwise 0. Hamming distance between two observations counts the number of not-matched measurements. For example,

	Gender	Employment Status
Kate	F	employed
John	M	unemployed
Adam	M	employed

The Hamming distance between Kate and John is 2 and between Kate and Adam is 1. One can also standardize the Hamming distance by dividing the number of categorical variables. That is

$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{p} \mathrm{I}(x_i \neq y_i)}{p}.$$

One special categorical variable is the *binary* variable which takes only two possible values: 0 (a certain attribute absent) or 1(a certain attribute present). A binary variable is called *asymmetric* if one of the two states (e.g. state "0") is interpreted as more informative than the other state. For example, Married (1) or Not Married (0); not married can be single, divorced or widowed. If both observations have value "0", we can not say if they are the same or different. Therefore, for asymmetric binary variable, the distance can be calculated as

$$d(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{p} \mathrm{I}(x_i \neq y_i)}{p - \#\text{ of 0-0 pairs}}$$

this is also called the *Jaccard coefficient.*

2.3 Distance for Mixed Variable Types

When the multivariate measurements are mixed of quantitative and categorical data, *Gower's coefficient* can be calculated to measure the distance between two observations:

$$d(\mathbf{x_i}, \mathbf{x_j}) = \frac{\sum_{k=1}^{p} \delta_{ijk} d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

where

$$\delta_{ijk} = \begin{cases} 1 & \text{if we could use the variable } k \text{ to compare observations } i \text{ and } j, \\ 0 & \text{if we could not tell whether observations } i \text{ and } j \text{ are the same or not using variable } k. \end{cases}$$

and

$$d_{ijk} = \begin{cases} \frac{|x_{ik} - x_{jk}|}{\text{range of variable } k} & \text{for quantitative variables,} \\ \mathrm{I}(x_{ij} \neq x_{jk}) & \text{for categorical variables.} \end{cases}$$

Example: Find the Gower's coefficients for the following three individuals:

	Gender	Hair Color	Asian	Height	Weight
Kate	F	Brown	Yes	60	80
John	M	Grey	No	50	60
Adam	M	Brown	No	70	90

Among the categorical variables "Gender", "Hair color", "Asian", "Asian" is asymmetric binary. We can construct a working table to find the distance between the observations.

k	δ_{ijk}	d_{ijk}	$\delta_{ijk} \times d_{ijk}$
1	1	1	1
2	1	1	1
3	1	1	1
4	1	$\frac{10}{20}$	0.5
5	1	$\frac{20}{30}$	$\frac{2}{3}$

k	δ_{ijk}	d_{ijk}	$\delta_{ijk} \times d_{ijk}$
1			
2			
3			
4			
5			

k	δ_{ijk}	d_{ijk}	$\delta_{ijk} \times d_{ijk}$
1			
2			
3			
4			
5			

$$\begin{aligned} d(\text{Kate, John}) &= \frac{\sum_{k=1}^{p} \delta_{ijk} d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}} \\ &= \frac{4.167}{5} \\ &= 0.8334 \end{aligned}$$

$$\begin{aligned} d(\text{Kate, Adam}) &= \frac{\sum_{k=1}^{p} \delta_{ijk} d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}} \\ &= \underline{\hspace{2cm}} \\ &= \underline{\hspace{2cm}} \end{aligned}$$

$$\begin{aligned} d(\text{John, Adam}) &= \frac{\sum_{k=1}^{p} \delta_{ijk} d_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}} \\ &= \underline{\hspace{2cm}} \\ &= \underline{\hspace{2cm}} \end{aligned}$$

3 Multivariate Normal Distribution

Most of the inferential statistical methods, such as t tests, one-way ANOVA, are based on the normality assumption in univariate analysis. Similarly, some multivariate analysis techniques are based on the assumption that the data are from a *multivariate normal* distribution.

The univariate density function of normal distribution with mean μ and standard deviation σ is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \tag{1}$$

The term

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

in the exponent of the univariate normal density function measures the squared distance between x and μ in standard deviation units. This can be generalized in vectors for multivariate cases as

$$(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)$$

where μ and $\mathbf{\Sigma}$ are the mean vector and variance-covariance matrix. It can be shown that the p-dimensional multivariate normal density function is

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

As the constant $\frac{1}{\sqrt{2\pi}\sigma}$ is the normalizing constant such that the *area* under the density curve given in Equation (1) is 1, the constant $\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}$ is the normalizing constant ensuring that the *volume* under the surface defined in Equation (2) is 1.

3.1 Properties of Multivariate Normal Distribution

Suppose that the joint distribution of random variables $X_1, X_2, \cdots, X_p$ is a multivariate normal with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{\Sigma}$, then we have the following results.

- The marginal distribution of X_i is $N(\mu_i, \sqrt{\sigma_{ii}})$.
- For any pair of X_i, X_j, they are independent if and only if $Cov(X_i, X_j) = 0$.
- Any linear combination of X_i, $\mathbf{c}^T\mathbf{X} = c_1X_1 + c_2X_2 + \cdots + c_pX_p$ is distributed as $N(\mathbf{c}^T\boldsymbol{\mu}, \mathbf{c}^T\mathbf{\Sigma}\mathbf{c})$.
- All points with the same distance to the mean $\boldsymbol{\mu}$ form a *contour* of the multivariate normal distribution. Contours for the p-dimensional normal distribution are ellipsoids defined by $\mathbf{x}$ such that

  $$(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2.$$

  These ellipsoids are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, where λ_i and $\mathbf{e}_i$ are the eigenvalues and corresponding unit eigenvectors of the variance-covariance matrix $\mathbf{\Sigma}$.
- If $|\mathbf{\Sigma}| > 0$, that is covariance matrix $\mathbf{\Sigma}$ is positive definite, then $(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is distributed as χ^2_p, a chi-square with degrees of freedom p. The multivariate normal distribution $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ assigns probability $(1-\alpha)$ to the solid ellipsoid $\{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \le \chi^2_p(\alpha)\}$, where $\chi^2_p(\alpha)$ is the upper (100α)th percentile of the χ^2_p distribution.

3.2 Bivariate Normal Distribution

It can be shown that when $p = 2$, we have the bivariate normal density

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}\exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\} \tag{3}$$

where $\mu_1 = E(X_1), \mu_2 = E(X_2), \sigma_{11} = Var(X_1), \sigma_{22} = Var(X_2), \rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$

Example: Bivariate Normal Distribution
If $X_1 \sim N(\mu_1, \sigma_1), X_2 \sim N(\mu_2, \sigma_2)$, and they are independent.

(a) Find the joint density of X_1 and X_2.

(b) Find the joint density of X_1 and X_2 using the matrix form.

(c) How about if X_1 and X_2 are not independent?

Example: Surface and Contour Plot of Bivariate Normal Consider the bivariate normal distributions with the following mean vectors and covariance matrices:

$$\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \boldsymbol{\mu_3} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma_3} = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}.$$

The distributions of the three bivariate normal are shown in Figure 6 and their corresponding contour plots are shown in Figure 7. The values on the contour paths are the value of density $f(x_1, x_2)$ given in Equation (3).

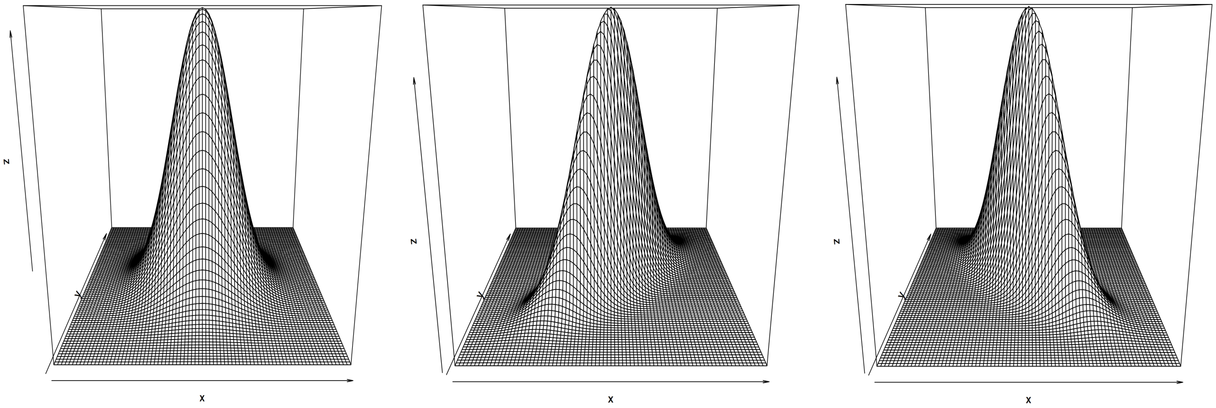

Figure 6: Density probability plots for three bivariate normal distributions

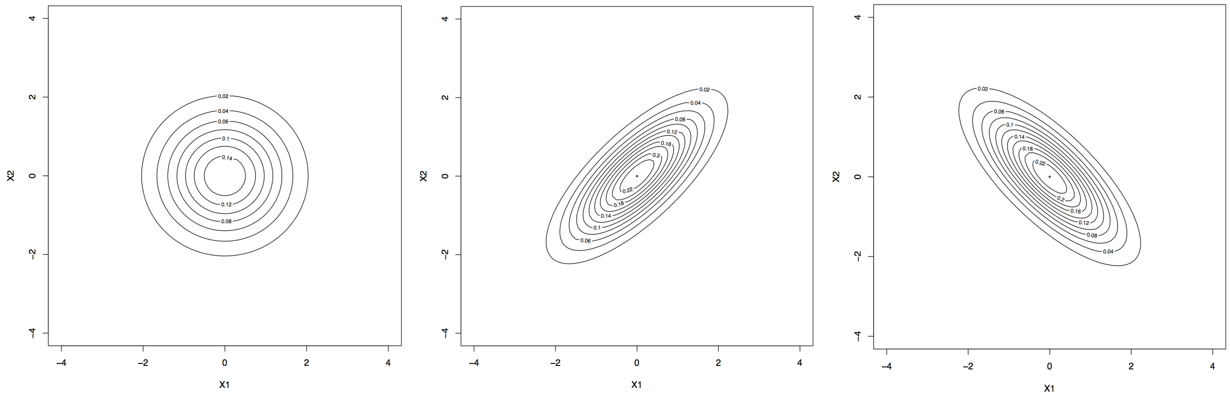

Figure 7: Contour plots of density for three bivariate normal distributions

3.3 Contour of Multivariate Normal Distribution

Definition

Contour at level c_0^2 is the collection of all points of $\mathbf{x}$ such that $f(\mathbf{x}) = c_0^2$. Since

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{(\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}{2}\right\} = c_0^2 \Longrightarrow (\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) = c^2$$

a $(1-\alpha) \times 100\%$ constant-density contour is the collection of points $\mathbf{x}$ such that

$$P((\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) \le c^2) = 1-\alpha.$$

Since $(\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) \sim \chi_p^2$, we get $c^2 = \chi_{p,\alpha}^2$. For example, a 95% contour for a bivariate normal distribution is the collection of all points $\mathbf{x}$ such that

$$(\mathbf{x}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) \le \chi_{2,0.05}^2 = 5.991$$

Example: Contour of Multivariate Normal Distribution

- Suppose $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, find a 95% contour for the bivariate normal distribution.
  For bivariate normal distribution, a 95% contour is

  $$(\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mu) = 5.991 \Longrightarrow \begin{bmatrix} x_1, x_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2 = 5.991$$

  which is a circle with radius $r = \sqrt{5.991}$.

- Suppose $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$, find a 95% contour for the bivariate normal distribution.
  For bivariate normal distribution, a 95% contour is

  $$(\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mu) = 5.991 \Longrightarrow$$

- Suppose $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{3}{4} \\ \frac{3}{4} & 1 \end{bmatrix}$, find a 95% contour for the bivariate normal distribution.
  For bivariate normal distribution, a 95% contour is

  $$(\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mu) = 5.991 \Longrightarrow \begin{bmatrix} x_1, x_2 \end{bmatrix} \frac{\begin{bmatrix} 1 & -\frac{3}{4} \\ -\frac{3}{4} & 1 \end{bmatrix}}{1 - \frac{9}{16}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{16}{7 \times 5.991} x_1^2 - \frac{24}{7 \times 5.991} x_1 x_2 + \frac{16}{7 \times 5.991} x_2^2 = 1$$

  which is an ellipse center $(0, 0)$ with axes $\pm \frac{\sqrt{5.991 \times 7}}{2}$ and $\pm \frac{\sqrt{5.991}}{2}$ and rotating counterclockwise 45 degree.

  Side-note: The generate form of a rotated counterclockwise about the origin through an angle θ with axes $\pm a$ and $\pm b$ is

  $$x^2 \left(\frac{\cos^2 \theta}{a^2} + \frac{\sin^2 \theta}{b^2} \right) + 2xy \left(\frac{\cos \theta \sin \theta}{a^2} - \frac{\cos \theta \sin \theta}{b^2} \right) + y^2 \left(\frac{\sin^2 \theta}{a^2} + \frac{\cos^2 \theta}{b^2} \right) = 1$$

To find the axes a^2, b^2 and the angle θ, solve for equations

$$\begin{cases} \frac{\cos^2\theta}{a^2} + \frac{\sin^2\theta}{b^2} = \frac{16}{7\times 5.991} \\ \frac{\sin^2\theta}{a^2} + \frac{\cos^2\theta}{b^2} = \frac{16}{7\times 5.991} \\ \frac{\cos\theta\sin\theta}{a^2} - \frac{\cos\theta\sin\theta}{b^2} = -\frac{12}{7\times 5.991} \end{cases}$$

The solutions are $\theta = 45, a^2 = \frac{5.991\times 7}{4}, b^2 = \frac{5.991}{4}$ or $\theta = 135, a^2 = \frac{5.991}{4}, b^2 = \frac{5.991\times 7}{4}$. They are the same ellipse.

Theorem
The contour $(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \le c^2$ is an ellipsoids centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, where λ_i and $\mathbf{e}_i$ are the eigenvalues and corresponding unit eigenvectors of the variance-covariance matrix $\mathbf{\Sigma}$.

Revisit the examples applying the theorem.

Example: find contour for the bivariate normal distribution by eigen-pairs of $\mathbf{\Sigma}$

- Suppose $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$, find a 95% contour for the bivariate normal distribution.
  The eigenvalues of $\mathbf{\Sigma}$ are

  $$|\mathbf{\Sigma} - \lambda\mathbf{I}| = \begin{vmatrix} 1-\lambda & 0 \\ 0 & 4-\lambda \end{vmatrix} = (1-\lambda)(4-\lambda) = 0$$

  which gives the eigenvalues are $\lambda_1 = 4$ and $\lambda_2 = 1$. For $\lambda_1 = 4$, we have

  $$\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 4\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

  which gives $x_1 = 4x_1, 4x_2 = 4x_2$; therefore the eigenvector could be $\mathbf{e_1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ For $\lambda_2 = 1$, the eigenvector could be $\mathbf{e_2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The axes are $\pm c\sqrt{\lambda_1}\mathbf{e}_1 = \pm\sqrt{5.991}\sqrt{4}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \pm\begin{bmatrix} 0 \\ 4.895 \end{bmatrix}$ and $\pm c\sqrt{\lambda_2}\mathbf{e}_2 = \pm\sqrt{5.991}\sqrt{1}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \pm\begin{bmatrix} 2.448 \\ 0 \end{bmatrix}$.

- Suppose $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & \frac{3}{4} \\ \frac{3}{4} & 1 \end{bmatrix}$, find a 95% contour for the bivariate normal distribution.

4 The Sampling Distribution of $\bar{\mathbf{X}}$ and S

In the univariate case, sample mean $\bar{X}$ is an unbiased estimator of the population mean μ. Inferences on μ is based on the sampling distribution of $\bar{X}$. It is known that $E(\bar{X}) = \mu, Var(\bar{X}) = \frac{\sigma^2}{n}$, where σ^2 is the population variance. If population distribution is normal, $\bar{X}$ is also normally distributed; if the population distribution is non-normal, by central limit theorem, when the sample size is large enough, $\bar{X}$ is approximately normally distributed. As for the sample variance s^2, we have $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. And $\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} \sim \chi^2_n$. Before generalizing to multivariate cases, let's review distributions related to normal for univariate variables.

4.1 Distributions Related to Normal Distribution

Most of the following conclusions can be shown by using moment-generating function.

- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.
- If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$, a chi-square distribution with degrees of freedom $df = 1$.
- If $W \sim \chi^2_p, V \sim \chi^2_q$, and they are independent, then
$$W + V \sim \chi^2_{p+q} \text{ and } \quad \frac{W/p}{V/q} \sim F_{p,q} \text{ (an F distribution with } df_n = p, df_d = q).$$
- If $Z \sim N(0,1), W \sim \chi^2_\gamma$, and they are independent, then
$$\frac{Z}{\sqrt{\frac{W}{\gamma}}} \sim t_\gamma \quad \text{(a } t \text{ distribution with degrees of freedom } df = \gamma)$$

4.2 Applications to Distributions Related to Sample Means

Suppose that $X_1, X_2, \cdots, X_n$ are iid from a normal with mean μ and variance σ^2, then the sample mean is defined as $\bar{X} = \frac{\sum X_i}{n}$. Then we have

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- $\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$ follows a chi-square distribution with $df = n$, denoted as χ^2_n.
- $\displaystyle\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2}$ follows a chi-square distribution with $df = n - 1$, denoted as χ^2_{n-1}. It can be shown that S^2 and $\bar{X}$ are independent.
- $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{[\frac{(n-1)S^2}{\sigma^2}]/(n-1)}} \sim t_{n-1}$. Note that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, therefore
$$T^2 = \left(\frac{\bar{X} - \mu}{S/\sqrt{n}}\right)^2 = \frac{\frac{n(\bar{X}-\mu)^2}{\sigma^2}/1}{\frac{(n-1)S^2}{\sigma^2}/(n-1)} \sim F_{1,n-1}$$

4.3 Generalize to Multivariate Cases

Univariate Cases	**Multivariate Cases**
$\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$	$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma})$
$n(\bar{X} - \mu)(\sigma^2)^{-1}(\bar{X} - \mu) \xrightarrow{D} \chi^2_1$	$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$
$n(\bar{X}-\mu)(S^2)^{-1}(\bar{X}-\mu) \xrightarrow{D} \chi^2_1$ when n is large enough	$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$ when $(n-p)$ is large.

5 Review Exercises

1. We consider the modified admission data set which contain 400 graduate school admissions decisions. There are five variables: admit/don?t admit to graduate school (1=admitted and 0=not admitted), GRE (Graduate Record Exam scores), GPA (grade point average), prestige of the undergraduate institution (values from 1 to 4 with 1 having the highest prestige), whether the undergraduate institution is in Ontario or not. Use the most proper distance metric to find the distance between the following three individuals:

Admit	GRE	GPA	Prestige	Ontario
0	380	3.61	3	Yes
1	660	3.67	3	NO
1	800	4.00	1	NO

   Suppose the range of GRE is 580 and the range of GPA is 1.74.

2. Consider a bivariate normal population with $\mu_1 = 1, \mu_2 = 3, \sigma_{11} = 2, \sigma_{22} = 1, \rho_{12} = -0.8$.
   (a) Write out the bivariate normal density.
   (b) Write out the squared Mahalanobis distance $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ as a function of x_1 and x_2.
   (c) Determine and sketch the constant -density contour that contains 90% of the probability.

Revisit the Learning Outcomes

After finishing this chapter, students should be able to

- Plot the index scatter plots matrix, draw a stars plot and Chernoff faces plot of a given multivariate data set using R, and interpret the plots.
- Verify whether a distance function is valid.
- Choose the proper distance metric to calculate the distance between observations based on the data types for multivariate data.
- Write down the density function of a multivariate normal distribution.
- Describe the properties of a multivariate normal distribution.
- Find a $(1 - \alpha) \times 100\%$ contour for a given bivariate normal distribution.
- Explain the distributions related to the sample mean vector and sample covariance matrix.