Chapter 2

Numeric Attributes

In this chapter, we discuss basic statistical methods for exploratory data analysis of numeric attributes. We look at measures of central tendency or location, measures of dispersion, and measures of linear dependence or association between attributes.

2.1 Univariate Analysis

Univariate analysis focuses on a single attribute at a time, thus the data matrix $\mathbf{D}$ can be thought of as a $n \times 1$ matrix, or simply a column vector, given as

$$\mathbf{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where X is the numeric attribute of interest, with $x_i \in \mathbb{R}$.

X is assumed to be a random variable, with each point x_i $(1 \leq i \leq n)$ itself treated as an identity random variable. We also treat the variables x_i as being independent and identically distributed as X, i.e., we assume that the observed data is a random sample drawn from X. In the vector view, we treat the sample as an n-dimensional vector, and write $X \in \mathbb{R}^n$.

Typically in data analysis, the probability density or mass function $f(x)$ and the cumulative distribution function $F(x)$, for attribute X, are both unknown. However, we can estimate these distributions directly from the data sample, which also allows us to compute statistics to estimate several important population parameters.

Empirical Cumulative Distribution Function The *empirical cumulative distribution function* of X is given as

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \tag{2.1}$$

where

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

is a binary *indicator variable*, that indicates whether the given condition is satisfied or not. Intuitively, to obtain the empirical CDF we compute for each value $x \in \mathbb{R}$, how many points in the sample are less than or equal to x. The empirical CDF puts a probability mass of $\frac{1}{n}$ at each point x_i.

Inverse Cumulative Distribution Function Define the *inverse cumulative distribution function* or *quantile function* for a random variable X as follows

$$F^{-1}(q) = \min\{x : F(x) > q\} \qquad \text{for } q \in [0, 1] \tag{2.2}$$

That is, the inverse CDF gives the least value of X, for which q fraction of the values are higher, and $1 - q$ fraction of the values are lower. The *empirical inverse cumulative distribution function* $\hat{F}^{-1}$ can be obtained from (2.1).

Empirical Probability Mass Function The *empirical probability mass function* of X is given as

$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x) \tag{2.3}$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

The empirical PMF also puts a probability mass of $\frac{1}{n}$ at each point x_i.

2.1.1 Measures of Central Tendency

Mean

The *mean* or *expectation* or *expected value* of a random variable X is the arithmetic average of the values of X. It provides a one-number summary of the *location* or *central tendency* for the distribution of X.

The mean or expected value of a discrete random variable X is defined as

$$\mu = E[X] = \sum_x x \; f(x) \tag{2.4}$$

where $f(x)$ is the probability mass function of X.

The expected value of a continuous random variable X is defined as

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{2.5}$$

where $f(x)$ is the probability density function of X.

Sample Mean The *sample mean* is a statistic, i.e., a function $\hat{\mu} : \{x_1, x_2, \cdots, x_n\} \to \mathbb{R}$, defined as the average value of x_i's

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.6}$$

It serves as an estimator for the unknown mean value μ of X. It can be intuitively derived by plugging in the empirical PMF $\hat{f}(x)$ in (2.4)

$$\hat{\mu} = \sum_x x \hat{f}(x) = \sum_x x \left(\frac{1}{n} \sum_{i=1}^{n} I(x_i = x) \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample Mean is Unbiased An estimator $\hat{\theta}$ is called an *unbiased estimator* for parameter θ if $E[\hat{\theta}] = \theta$ for every possible value of θ. The sample mean $\hat{\mu}$ is an unbiased estimator for the population mean μ, since

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[x_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu \tag{2.7}$$

Above we used the fact that the random variables x_i are IID according to X, which implies that they have the same mean μ as X, i.e., $E[x_i] = \mu$ for all x_i. We also used the fact that E is a *linear operator*, i.e., for any two random variables X and Y, and real numbers a and b, we have $E[aX + bY] = aE[X] + bE[Y]$.

Robustness We say that a statistic is *robust* if it is not affected by extreme values (such as outliers) in the data. The sample mean is unfortunately not robust, since a single large value can skew the average. A more robust measure is the *trimmed mean* obtained after discarding a small fraction of extreme values on one or both ends. Furthermore, the mean can be somewhat misleading in that it is typically not a value that occurs in the sample, and it may not even be a value that the

random variable can actually assume (for a discrete random variable). For example, the number of cars per capita is an integer valued random variable, but according to the US Bureau of Transportation Studies, the average number of passenger cars in the US was 0.45 in 2008 (137.1 million cars, with a population size of 304.4 million). Obviously, one cannot own 0.45 cars; it simply means that on average there are 45 cars per 100 people.

Median

The *median* of a random variable is defined as the value m, such that

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

In other words, the median m is the "middle-most" value; half of the values of X are less and half of the values of X are more than m. In terms of the (inverse) cumulative distribution function, the median is therefore the value m for which

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5) \tag{2.8}$$

The *sample median* can be obtained from the empirical CDF (2.1) or the empirical inverse CDF (2.2) by computing

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5) \tag{2.9}$$

A simpler approach to compute the sample median is to first sort all the values x_i ($i \in [1, n]$) in increasing order. If n is odd, the median is the value at position $\frac{n+1}{2}$. If n is even, the values at positions $\frac{n}{2}$ and $\frac{n+2}{2}$ are both medians.

Unlike the mean, median is robust, since it is not affected very much by extreme values. Also, it is a value that occurs in the sample, and a value the random variable can actually assume.

Mode

The *mode* of a random variable X is the value at which the probability mass function or the probability density function attains its maximum value, depending on whether X is discrete or continuous, respectively.

The *sample mode* is a value for which the empirical probability function (2.3) attains its maximum, given as

$$\text{mode}(X) = \arg\max_x \hat{f}(x) \tag{2.10}$$

The mode may not be a very useful measure of central tendency for a sample, since by chance an unrepresentative element may be the most frequent element. Furthermore, if all values in the sample are distinct, each of them will be the mode.

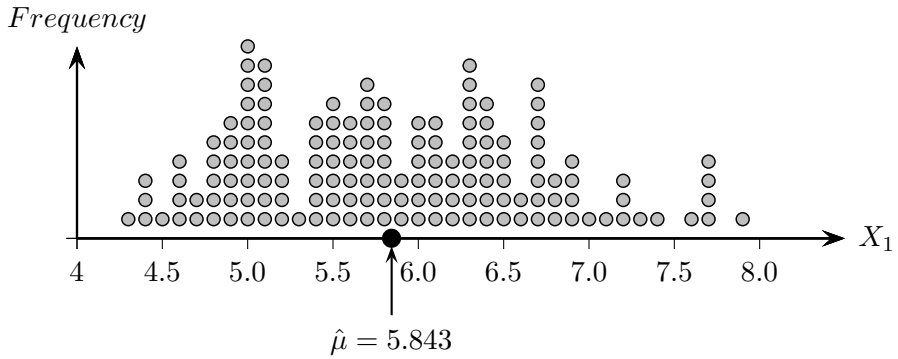

Figure 2.1: Sample Mean for `sepal length`. Multiple occurrences of the same value are shown stacked.

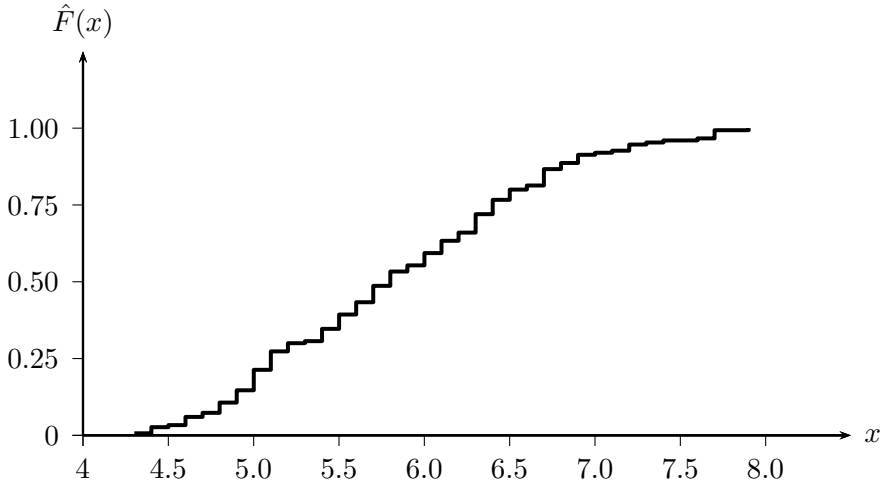

Figure 2.2: Empirical CDF: `sepal length`

Example 2.1 (Sample Mean, Median, and Mode): Consider attribute `sepal length` (X_1) in the Iris dataset, whose values are shown in Table 1.2. The sample mean is given as follows

$$\hat{\mu} = \frac{1}{150}(5.9 + 6.9 + \cdots + 7.7 + 5.1) = \frac{876.5}{150} = 5.843$$

Figure 2.1 shows all 150 values of `sepal length`, and the sample mean. Figure 2.2 shows the empirical CDF and Figure 2.3 shows the empirical inverse CDF for `sepal length`.

Since $n = 150$ is even, the sample median is the value at positions $\frac{n}{2} = 75$ and $\frac{n+2}{2} = 76$ in sorted order. For `sepal length` both these values are 5.8, thus the sample median is 5.8. From the inverse CDF in Figure 2.3, we can see that

$$\hat{F}(5.8) = 0.5 \text{ or } 5.8 = \hat{F}^{-1}(0.5)$$

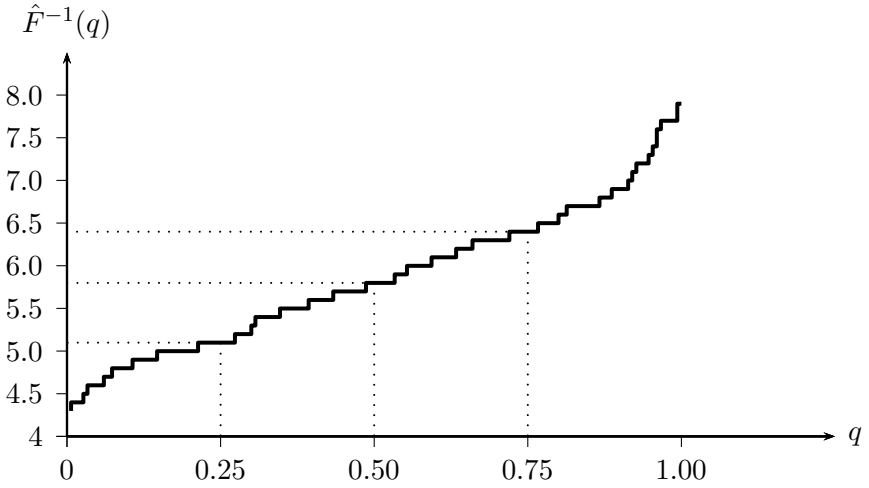

Figure 2.3: Empirical Inverse CDF: `sepal length`

The sample mode for `sepal length` is 5, which can be observed from the frequency of 5 in Figure 2.1. The empirical probability mass at $x = 5$ is

$$\hat{f}(5) = \frac{10}{150} = 0.067$$

2.1.2 Measures of Dispersion

The measures of dispersion give an indication about the spread or variation in the values of a random variable.

Range

The *value range* or simply *range* of a random variable X is the difference between the maximum and minimum values of X, given as

$$r = \max\{X\} - \min\{X\} \tag{2.11}$$

The (value) range of X is a population parameter, not to be confused with the range of the function X, which is the set of all the values X can assume. Which range is being used should be clear from the context.

The *sample range* is a statistic, given as

$$\hat{r} = \max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\} \tag{2.12}$$

By definition, range is sensitive to extreme values, and thus is not robust.

Inter-Quartile Range

Quartiles are special values of the quantile function (2.2), that divide the data into 4 equal parts. That is, quartiles correspond to the quantile values of 0.25, 0.5, 0.75, and 1.0. The *first quartile* is the value $q_1 = F^{-1}(0.25)$, to the left of which 25% of the points lie, the *second quartile* is the same as the median value $q_2 = F^{-1}(0.5)$, to the left of which 50% of the points lie, the third quartile $q_3 = F^{-1}(0.75)$ is the value to the left of which 75% of the points lie, and the fourth quartile is the maximum value of X, to the left of which 100% of the points lie.

A more robust measure of the dispersion of X is the *inter-quartile range (IQR)*, defined as

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25) \tag{2.13}$$

IQR can also be thought of as a *trimmed range*, where we discard 25% of the low and high values of X. Or put differently, it is the range for the middle 50% of the values of X. IQR is robust by definition.

The *sample IQR* can be obtained by plugging in the empirical inverse CDF in (2.13)

$$\widehat{IQR} = \hat{q}_3 - \hat{q}_1 = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25) \tag{2.14}$$

Variance and Standard Deviation

The *variance* of a random variable X provides a measure of how much the values of X deviate from the mean or expected value of X. More formally, variance is the expected value of the squared deviation from the mean, defined as

$$\sigma^2 = var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx & \text{if } X \text{ is continuous} \end{cases} \tag{2.15}$$

The *standard deviation*, σ, is defined as the positive square root of the variance, σ^2.

We can also write the variance as the difference between the expectation of X^2 and the square of the expectation of X

$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned} \tag{2.16}$$

It is worth noting that variance is in fact the *second moment about the mean*, corresponding to $r = 2$, which is a special case of the *r-th moment about the mean*

for a random variable X, defined as

$$E\left[(\mathbf{x}-\mu)^r\right]$$

Sample Variance The *sample variance* is defined as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{2.17}$$

It is the average squared deviation of the data values x_i from the sample mean $\hat{\mu}$, and can be derived by plugging in the empirical probability function $\hat{f}$ from (2.3) into (2.15), since

$$\hat{\sigma}^2 = \sum_x (x-\hat{\mu})^2\hat{f}(x) = \sum_x (x-\hat{\mu})^2\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i=x)\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

The *sample standard deviation* is given as the square root of the sample variance

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2} \tag{2.18}$$

The *standard score*, also called the *z-score* , of a sample value x_i is the number of standard deviations away the value is from the mean

$$z_i = \frac{x_i-\hat{\mu}}{\hat{\sigma}} \tag{2.19}$$

Put differently, the z-score of x_i measures the deviation of x_i from the mean value $\hat{\mu}$, in units of $\hat{\sigma}$.

Geometric Interpretation of Sample Variance We can treat the data sample for attribute X as a vector in n-dimensional space, where n is the sample size. That is, we write $X = (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^n$. Further, let

$$Z = X - \mathbf{1}\cdot\hat{\mu} = \begin{pmatrix} x_1-\hat{\mu} \\ x_2-\hat{\mu} \\ \vdots \\ x_n-\hat{\mu}\end{pmatrix}$$

denote the mean subtracted attribute vector, where $\mathbf{1}\in\mathbb{R}^n$ is the n-dimensional vector all of whose elements have value 1. We can rewrite (2.17) in terms of the magnitude of Z, i.e., the dot product of Z with itself

$$\hat{\sigma}^2 = \frac{1}{n}\|Z\|^2 = \frac{1}{n}Z^TZ = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{2.20}$$

The sample variance can thus be interpreted as the squared magnitude of the centered attribute vector, or the dot product of the centered attribute vector with itself, normalized by the sample size.

Variance of the Sample Mean Since the sample mean $\hat{\mu}$ is itself a statistic, we can compute its mean value and variance. The expected value of the sample mean is simply μ, as we saw in (2.7). To derive an expression for the variance of the sample mean, we utilize the fact that since the random variables x_i are all independent, and thus

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i)$$

Further since all the x_i's are identically distributed as X, they have the same variance as X, i.e.,

$$var(x_i) = \sigma^2 \text{ for all } i$$

Combining the above two facts, we get

$$var\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2 \tag{2.21}$$

Further, note that

$$E\left[\sum_{i=1}^{n} x_i\right] = n\mu \tag{2.22}$$

Using (2.16), (2.21), and (2.22), the variance of the sample mean $\hat{\mu}$ can be computed as

$$\begin{aligned} var(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = E[\hat{\mu}^2] - \mu^2 &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \frac{1}{n^2}E\left[\sum_{i=1}^{n} x_i\right]^2 \\ &= \frac{1}{n^2}\left(E\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] - E\left[\sum_{i=1}^{n} x_i\right]^2\right) = \frac{1}{n^2}\, var\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{2.23}$$

In other words, the sample mean $\hat{\mu}$ varies or deviates from the mean μ in proportion to the population variance σ^2. However, the deviation can be made smaller by considering larger sample size n.

Sample Variance is Biased, but is Asymptotically Unbiased The sample variance in (2.17) is a *biased estimator* for the true population variance, σ^2, i.e., $E[\hat{\sigma}^2] \neq \sigma^2$. To show this we make use of the identity

$$\sum_{i=1}^{n}(x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{2.24}$$

Computing the expectation of $\hat{\sigma}^2$ by using (2.24) in the first step, we get

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right] - E[(\hat{\mu} - \mu)^2] \tag{2.25}$$

Recall that the random variables x_i are IID according to X, which means that they have the same mean μ and variance σ^2 as X. This means that

$$E[(x_i - \mu)^2] = \sigma^2$$

Further, from (2.23) the sample mean $\hat{\mu}$ has variance $E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$. Plugging these into the (2.25) we get

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{1}{n}\, n\sigma^2 - \frac{\sigma^2}{n} \\ &= \left(\frac{n-1}{n}\right)\sigma^2 \end{aligned} \tag{2.26}$$

In other words, $\hat{\sigma}^2$ is a biased estimator of σ^2, since its expected value differs from the population variance by a factor of $\frac{n-1}{n}$. However, it is *asymptotically unbiased*, that is, the bias vanishes as $n \to \infty$, since

$$\lim_{n\to\infty} \frac{n-1}{n} = \lim_{n\to\infty} 1 - \frac{1}{n} = 1$$

Put another way, as the sample size increases, we have

$$E[\hat{\sigma}^2] \to \sigma^2 \qquad \text{as } n \to \infty$$

Example 2.2: Consider the data sample for `sepal length` shown in Figure 2.1. We can see that the sample range is given as

$$max_i\{x_i\} - min_i\{x_i\} = 7.9 - 4.3 = 3.6$$

From the inverse CDF for `sepal length` in Figure 2.3, we can find the sample IQR as follows

$$\begin{aligned} \hat{q}_1 &= \hat{F}^{-1}(0.25) = 5.1 \\ \hat{q}_3 &= \hat{F}^{-1}(0.75) = 6.4 \\ \widehat{IQR} &= \hat{q}_3 - \hat{q}_1 = 6.4 - 5.1 = 1.3 \end{aligned}$$

The sample variance can be computed from the centered data vector via the expression (2.20)

$$\hat{\sigma}^2 = \frac{1}{n}(X - \mathbf{1}\cdot\hat{\mu})^T(X - \mathbf{1}\cdot\hat{\mu}) = 102.168/150 = 0.681$$

The sample standard deviation is then

$$\hat{\sigma} = \sqrt{0.681} = 0.825$$

2.2 Bivariate Analysis

In bivariate analysis, we consider two attributes at the same time. We are specifically interested in understanding the association or dependence between them, if any. We thus restrict our attention to the two numeric attributes of interest, X_1 and X_2, with the data $\mathbf{D}$ represented as a $n \times 2$ matrix

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \tag{2.27}$$

Geometrically, we can think of $\mathbf{D}$ in two ways. It can be viewed as n points or vectors in two dimensional space over the attributes X_1 and X_2, i.e., $\mathbf{x}_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$. Alternatively, it can be viewed as two points or vectors in an n-dimensional space comprising the points, i.e., each column is a vector in $\mathbb{R}^n$, as follows

$$X_1 = (x_{11}, x_{21}, \cdots, x_{n1})^T$$
$$X_2 = (x_{12}, x_{22}, \cdots, x_{n2})^T$$

In the probabilistic view, the column vector $\mathbf{X} = (X_1, X_2)^T$ is considered a bivariate vector random variable, and the points $\mathbf{x}_i$ $(1 \leq i \leq n)$ are treated as a random sample drawn from $\mathbf{X}$, i.e., $\mathbf{x}_i$'s are treated as independent and identically distributed as $\mathbf{X}$.

Empirical Joint Probability Mass Function The *empirical joint probability mass function* for $\mathbf{X}$ is given as

$$\hat{f}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \tag{2.28}$$

$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^{n} I(x_{i1} = x_1, x_{i2} = x_2)$$

where I is a indicator variable which is true only when the condition holds

$$I(\mathbf{x}_i = \mathbf{x}) = \begin{cases} 1 & \text{if } x_{i1} = x_1 \text{ and } x_{i2} = x_2 \\ 0 & \text{otherwise} \end{cases}$$

As in the univariate case, the probability function puts a probability mass of $\frac{1}{n}$ at each point in the data sample.

2.2.1 Measures of Location and Dispersion

Mean The bivariate mean is defined as the expected value of the vector random variable $\mathbf{X}$, defined as follows

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{2.29}$$

In other words, the bivariate mean vector is simply the vector of expected values along each attribute.

The sample mean vector can be obtained from $\hat{f}_{X_1}$ and $\hat{f}_{X_2}$, the empirical probability mass functions of X_1 and X_2, respectively, computed via (2.6). It can also be computed from the joint empirical PMF in (2.2)

$$\hat{\boldsymbol{\mu}} = \sum_{\mathbf{x}} \mathbf{x}\hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x}\left(\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x})\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \tag{2.30}$$

Variance We can compute the variance along each attribute, namely σ_1^2 for X_1 and σ_2^2 for X_2 using (2.15). The *total variance* (1.9) is given as

$$\sigma_1^2 + \sigma_2^2$$

The sample variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ can be estimated using (2.17), and the *sample total variance* is simply $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$.

2.2.2 Measures of Association

Covariance The *covariance* between two attributes X_1 and X_2 provides a measure of the association or linear dependence between them, and is defined as

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] \tag{2.31}$$

By linearity of expectation, we have

$$\begin{aligned} \sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] \\ &= E[X_1X_2] - \mu_2E[X_1] - \mu_1E[X_2] + \mu_1\mu_2 \\ &= E[X_1X_2] - \mu_1\mu_2 \\ &= E[X_1X_2] - E[X_1]E[X_2] \end{aligned} \tag{2.32}$$

The expression above can be seen as a generalization of (2.16) to the two dimensional case.

If X_1 and X_2 are independent random variables, then we conclude that their covariance is zero. This is because if X_1 and X_2 are independent, then we have

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

which in turn implies that

$$\sigma_{12} = 0$$

However, the converse is not true. That is, if $\sigma_{12} = 0$, one cannot claim that X_1 and X_2 are independent. All we can say is that there is no linear dependence between them, but we cannot rule out that there might be a higher order relationship or dependence between the two attributes.

The *sample covariance* between X_1 and X_2 is given as

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2) \tag{2.33}$$

It can be derived by substituting the empirical joint probability mass function $\hat{f}(x_1, x_2)$ from (2.2) into (2.31), as follows

$$\begin{aligned}
\hat{\sigma}_{12} &= E[(X_1 - \hat{\mu}_1)(X_2 - \hat{\mu}_2)] \\
&= \sum_{\mathbf{x}=(x_1,x_2)^T} (x_1 - \hat{\mu}_1)(x_2 - \hat{\mu}_2)\hat{f}(x_1, x_2) \\
&= \frac{1}{n} \sum_{\mathbf{x}=(x_1,x_2)^T} \sum_{i=1}^{n} (x_1 - \hat{\mu}_1) \cdot (x_2 - \hat{\mu}_2) \cdot I(x_{i1} = x_1, x_{i2} = x_2) \\
&= \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)
\end{aligned}$$

Notice that sample covariance is a generalization of the sample variance (2.17), since

$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_1)(x_i - \mu_1) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_1)^2 = \hat{\sigma}_1^2$$

Correlation The *correlation* between variables X_1 and X_2 is the *standardized covariance*, obtained by normalizing the covariance with the standard deviation of each variable, given as

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}} \tag{2.34}$$

The *sample correlation* for attributes X_1 and X_2 is given as

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}} \tag{2.35}$$

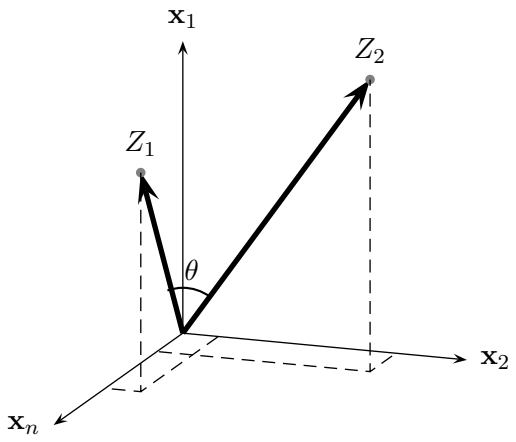

Figure 2.4: Geometric Interpretation of Covariance and Correlation. The two centered attribute vectors are shown in the (conceptual) n-dimensional space $\mathbb{R}^n$ spanned by the n points.

Geometric Interpretation of Sample Covariance and Correlation Let Z_1 and Z_2 denote the centered attribute vectors in $\mathbb{R}^n$, given as follows

$$Z_1 = X_1 - \mathbf{1} \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \qquad Z_2 = X_2 - \mathbf{1} \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix} \tag{2.36}$$

The sample covariance (2.33) can then be written as

$$\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n} \tag{2.37}$$

In other words, the covariance between the two attributes is simply the dot product between the two centered attribute vectors, normalized by the sample size. The above can be seen as a generalization of the sample variance given in (2.20).

The sample correlation (2.35) can be written as

$$\hat{\rho}_{12} = \frac{Z_1^T Z_2}{\sqrt{Z_1^T Z_1}\sqrt{Z_2^T Z_2}} = \frac{Z_1^T Z_2}{\|Z_1\| \ \|Z_2\|} = \left(\frac{Z_1}{\|Z_1\|}\right)^T \left(\frac{Z_2}{\|Z_2\|}\right) = \cos\theta \tag{2.38}$$

Thus the correlation coefficient is simply the cosine of the angle (see (1.6)) between the two centered attribute vectors, as illustrated in Figure 2.4.

Covariance Matrix The variance-covariance information for the two attributes X_1 and X_2 can be summarized in the square 2×2 *covariance matrix* , given as

$$\begin{aligned}\boldsymbol{\Sigma} &= E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \\ &= E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} \begin{pmatrix} X_1 - \mu_1 & X_2 - \mu_2 \end{pmatrix}\right] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \end{aligned} \tag{2.39}$$

Since $\sigma_{12} = \sigma_{21}$, $\boldsymbol{\Sigma}$ is a *symmetric* matrix. The covariance matrix records the attribute specific variances on the main diagonal, and the covariance information on the off-diagonal elements.

The *total variance* of the two attributes is given as the sum of the diagonal elements of $\boldsymbol{\Sigma}$, which is also called the *trace* of $\boldsymbol{\Sigma}$, given as

$$tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2$$

We immediately have $tr(\boldsymbol{\Sigma}) \geq 0$.

The *generalized variance* of the two attributes also considers the covariance, in addition to the attribute variances, and is given as the *determinant* of $\boldsymbol{\Sigma}$, $det(\boldsymbol{\Sigma})$. It is worth noting that the generalized covariance is non-negative, since

$$\det(\boldsymbol{\Sigma}) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2 = (1 - \rho_{12}^2)\sigma_1^2\sigma_2^2 \tag{2.40}$$

Note that $|\rho_{12}| \leq 1$, implies that $\rho_{12}^2 \leq 1$, which in turn implies that $\det(\boldsymbol{\Sigma}) \geq 0$, i.e., the determinant is non-negative.

The *sample covariance matrix* is given as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix} \tag{2.41}$$

$\widehat{\boldsymbol{\Sigma}}$ shares the same properties as $\boldsymbol{\Sigma}$, and can be used to easily obtain the sample total and generalized variance.

Example 2.3 (Sample Mean and Covariance): Consider the `sepal length` and `sepal width` attributes for the Iris dataset, plotted in Figure 2.5. There are $n = 150$ points in the $d = 2$ dimensional attribute space. The sample mean vector is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

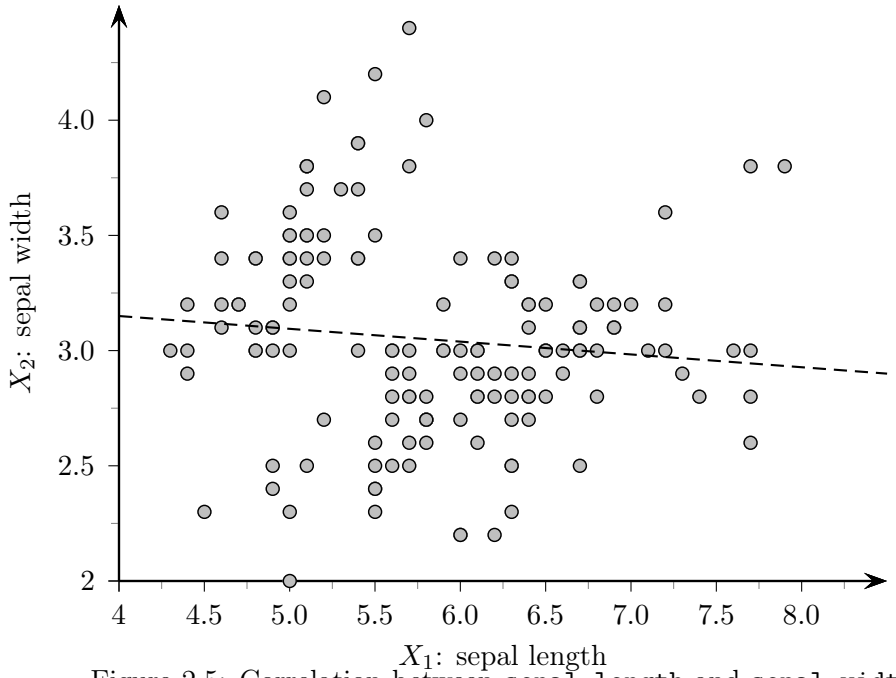

Figure 2.5: Correlation between `sepal length` and `sepal width`

The sample covariance matrix is given as

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The variance for `sepal length` is $\hat{\sigma}_1^2 = 0.681$, and that for `sepal width` is $\hat{\sigma}_2^2 = 0.187$. The covariance between the two attributes is $\hat{\sigma}_{12} = -0.039$, and the correlation between them is

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Thus there is a very weak negative correlation between these two attributes, as evidenced by the best linear fit line in Figure 2.5. Alternatively, we can consider the attributes `sepal length` and `sepal width` as two points in $\mathbb{R}^n$. The correlation is then the cosine of the angle between them; we have

$$\hat{\rho}_{12} = \cos\theta = -0.109, \text{ which implies that } \theta = \cos^{-1}(-0.109) = 96.26^\circ$$

The angle is close to 90°, i.e., the two attribute vectors are almost orthogonal, indicating weak correlation. Further, the angle being greater than 90° indicates negative correlation.

The sample total variance is given as

$$tr(\widehat{\mathbf{\Sigma}}) = 0.681 + 0.187 = 0.868$$

and the sample generalized variance is given as

$$det(\widehat{\mathbf{\Sigma}}) = 0.681 \cdot 0.187 - (-0.039)^2 = 0.126$$

2.2.3 Data Normalization

When analyzing two or more attributes it is often necessary to normalize the values of the attributes, especially in those cases where the values are vastly different in scale.

Range Normalization Let X be an attribute with values $x_1, x_2, \ldots, x_n$. In *range normalization*, each value is scaled by the range $\hat{r}$ of the attribute

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

After transformation the new attribute takes on values in the range $[0, 1]$.

Standard Normalization In *standard normalization*, each value is replaced by its z-score

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

After transformation, the new attribute has mean $\hat{\mu}' = 0$, and standard deviation $\hat{\sigma}' = 1$.

	Age (X_1)	Income (X_2)
$\mathbf{x}_1$	12	300
$\mathbf{x}_2$	14	500
$\mathbf{x}_3$	18	1000
$\mathbf{x}_4$	23	2000
$\mathbf{x}_5$	27	3500
$\mathbf{x}_6$	28	4000
$\mathbf{x}_7$	34	4300
$\mathbf{x}_8$	37	6000
$\mathbf{x}_9$	39	2500
$\mathbf{x}_{10}$	40	2700

Table 2.1: Dataset for Normalization

Example 2.4: Consider the example dataset shown in Table 2.1. The attributes `Age` and `Income` have very different scales, with the latter having much larger values. Consider the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$

$$\|\mathbf{x}_1 - \mathbf{x}_2\| = \left\|(2, 200)^T\right\| = \sqrt{2^2 + 200^2} = \sqrt{40004} = 200.01$$

As we can observe, the contribution of `Age` is overshadowed by the value of `Income`.

The range of `Age` is $\hat{r} = 40 - 12 = 28$, with the minimum value 12. After range normalization, the new attribute is given as

$$\texttt{Age}' = (0, 0.071, 0.214, 0.393, 0.536, 0.571, 0.786, 0.893, 0.964, 1)^T$$

For example, the value $x_{12} = 14$ is transformed into

$$x'_{12} = \frac{14 - 12}{28} = \frac{2}{28} = 0.071$$

Likewise `Income` is transformed into

$$\texttt{Income}' = (0, 0.035, 0.123, 0.298, 0.561, 0.649, 0.702, 1, 0.386, 0.421)^T$$

For z-normalization, we first compute the mean and standard deviation of both attributes

	Age	Income
μ	27.2	2680
σ	9.77	1726.15

`Age` is transformed into

$$\texttt{Age}' = (-1.56, -1.35, -0.94, -0.43, -0.02, 0.08, 0.70, 1.0, 1.21, 1.31)^T$$

For instance, the value $x_{12} = 14$ is transformed as

$$x'_{12} = \frac{14 - 27.2}{9.77} = -1.35$$

Likewise, `Income` is transformed as

$$\texttt{Income}' = (-1.38, -1.26, -0.97, -0.39, 0.48, 0.77, 0.94, 1.92, -0.10, 0.01)^T$$

2.3 Multivariate Analysis

In multivariate analysis, we consider all the d attributes $X_1, X_2, \cdots, X_d$. The full data is a $n \times d$ matrix, given as

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} \tag{2.42}$$

In the row view, the data can be considered as a set of n points or vectors in the d-dimensional attribute space

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T \in \mathbb{R}^d$$

In the column view, the data can be considered as a set of d points or vectors in the n-dimensional space spanned by the data points

$$X_j = (x_{1j}, x_{2j}, \cdots, x_{nj})^T \in \mathbb{R}^n$$

In the probabilistic view, the d attributes are modeled as a vector random variable, $\mathbf{X} = (X_1, X_2, \cdots, X_d)^T$, and the points $\mathbf{x}_i$ are considered to be a random sample drawn from $\mathbf{X}$, i.e., they are independent and identically distributed as $\mathbf{X}$.

Mean Generalizing (2.29), the *multivariate mean vector* is obtained by taking the mean of each attribute, given as

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix} \tag{2.43}$$

Generalizing (2.30), the *sample mean* is given as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{2.44}$$

Covariance Matrix Generalizing (2.39) to d-dimensions, the multivariate covariance information is captured by the $d \times d$ (square) symmetric *covariance matrix* that gives the covariance for each pair of attributes

$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix} \tag{2.45}$$

Each diagonal elements σ_i^2 specify the attribute variances, whereas the off-diagonal elements $\sigma_{ij} = \sigma_{ji}$ represent the covariance between attribute pairs X_i and X_j.

Covariance Matrix is Positive Semi-definite It is worth noting that $\mathbf{\Sigma}$ is a *positive semi-definite* matrix, i.e.,

$$\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} \geq 0 \text{ for any } d\text{-dimensional vector } \mathbf{a} \tag{2.46}$$

To see this, observe that

$$\begin{aligned}
\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} &= \mathbf{a}^T E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T]\mathbf{a} \\
&= E[\mathbf{a}^T(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{a}] \\
&= E[Y^2], \text{ where } Y \text{ is the random variable } Y = \mathbf{a}^T(\mathbf{X}-\boldsymbol{\mu}) = \sum_{i=1}^{d} a_i(X_i - \mu_i) \\
&\geq 0, \text{ since the expectation of a squared random variable is non-negative}
\end{aligned}$$

Since $\mathbf{\Sigma}$ is also symmetric, this implies that all the eigenvalues of $\mathbf{\Sigma}$ are real and non-negative. In other words the d eigenvalues of $\mathbf{\Sigma}$ can be arranged from the largest to the smallest as follows: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. A consequence is that the determinant of $\mathbf{\Sigma}$ is non-negative

$$det(\mathbf{\Sigma}) = \prod_{i=1}^{d} \lambda_i \geq 0 \tag{2.47}$$

Total and Generalized Variance The total variance is given as the trace of the covariance matrix

$$tr(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2 \tag{2.48}$$

Being a sum of squares, the total variance must be non-negative.

The generalized variance is defined as the determinant of the covariance matrix, $\det(\mathbf{\Sigma})$, also denoted as $|\mathbf{\Sigma}|$, gives a single value for the overall multivariate scatter. From (2.47) we have $det(\mathbf{\Sigma}) \geq 0$.

Sample Covariance Matrix The *sample covariance matrix* is given as

$$\widehat{\mathbf{\Sigma}} = E[(\mathbf{X}-\hat{\boldsymbol{\mu}})(\mathbf{X}-\hat{\boldsymbol{\mu}})^T] = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix} \tag{2.49}$$

Instead of computing the sample covariance matrix element-by-element, we can obtain it in a single matrix operation. Let $\mathbf{Z}$ represent the centered data matrix

(1.10), given as the matrix of centered attribute vectors $Z_i = X_i - \mathbf{1} \cdot \hat{\mu}_i$, where $\mathbf{1} \in \mathbb{R}^n$

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}^T = \begin{pmatrix} | & | & & | \\ Z_1 & Z_2 & \cdots & Z_d \\ | & | & & | \end{pmatrix}$$

Alternatively, the centered data matrix can also be written in terms of the centered points $\mathbf{z}_i = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}^T = \begin{pmatrix} \mathbf{x}_1^T - \hat{\boldsymbol{\mu}}^T \\ \mathbf{x}_2^T - \hat{\boldsymbol{\mu}}^T \\ \vdots \\ \mathbf{x}_n^T - \hat{\boldsymbol{\mu}}^T \end{pmatrix} = \begin{pmatrix} — & \mathbf{z}_1^T & — \\ — & \mathbf{z}_2^T & — \\ & \vdots & \\ — & \mathbf{z}_n^T & — \end{pmatrix}$$

In matrix notation, the sample covariance matrix can be written as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{Z}^T\mathbf{Z}\right) = \frac{1}{n}\begin{pmatrix} Z_1^T Z_1 & Z_1^T Z_2 & \cdots & Z_1^T Z_d \\ Z_2^T Z_1 & Z_2^T Z_2 & \cdots & Z_2^T Z_d \\ \vdots & \vdots & \ddots & \vdots \\ Z_d^T Z_1 & Z_d^T Z_2 & \cdots & Z_d^T Z_d \end{pmatrix} \tag{2.50}$$

The sample covariance matrix is thus given as the pair-wise *inner or dot products* of the centered attribute vectors, normalized by the sample size.

In terms of the centered points $\mathbf{z}_i$, the sample covariance matrix can also be written as a sum of rank-one matrices obtained as the *outer-product* of each centered point

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \cdot \mathbf{z}_i^T \tag{2.51}$$

Example 2.5 (Sample Mean and Covariance Matrix): Let us consider all four numeric attributes for the Iris dataset, namely `sepal length`, `sepal width`, `petal length`, and `petal width`. The multivariate sample mean vector is given as

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 & 3.054 & 3.759 & 1.199 \end{pmatrix}^T$$

and the sample covariance matrix is given as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.681 & -0.039 & 1.265 & 0.513 \\ -0.039 & 0.187 & -0.320 & -0.117 \\ 1.265 & -0.320 & 3.092 & 1.288 \\ 0.513 & -0.117 & 1.288 & 0.579 \end{pmatrix}$$

The sample total variance is given as

$$tr(\widehat{\mathbf{\Sigma}}) = 0.681 + 0.187 + 3.092 + 0.579 = 4.539$$

The generalized variance is given as

$$\det(\widehat{\mathbf{\Sigma}}) = 1.853 \times 10^{-3}$$

To illustrate the inner and outer product based computation of the covariance matrix, consider the 2-dimensional dataset

$$\mathbf{D} = \begin{pmatrix} A_1 & A_2 \\ \hline 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{pmatrix}$$

The mean vector is as follows

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} 15/3 \\ 8.7/3 \end{pmatrix} = \begin{pmatrix} 5 \\ 2.9 \end{pmatrix}$$

and the centered data matrix is then given as

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T = \begin{pmatrix} 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 5 & 2.9 \end{pmatrix} = \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$

The inner-product approach (2.50) to compute the sample covariance matrix gives

$$\begin{aligned} \widehat{\mathbf{\Sigma}} &= \frac{1}{3} \begin{pmatrix} -4 & 0 & 4 \\ -2.1 & -0.5 & 2.6 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix} \\ &= \frac{1}{3} \begin{pmatrix} 32 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix} \end{aligned}$$

Alternatively, the outer-product approach (2.51) gives

$$\begin{aligned} \widehat{\mathbf{\Sigma}} &= \frac{1}{3} \left[\begin{pmatrix} -4 \\ -2.1 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \end{pmatrix} + \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 & -0.5 \end{pmatrix} + \begin{pmatrix} 4 \\ 2.6 \end{pmatrix} \cdot \begin{pmatrix} 4 & 2.6 \end{pmatrix} \right] \\ &= \frac{1}{3} \left[\begin{pmatrix} 16.0 & 8.4 \\ 8.4 & 4.41 \end{pmatrix} + \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 0.25 \end{pmatrix} + \begin{pmatrix} 16.0 & 10.4 \\ 10.4 & 6.76 \end{pmatrix} \right] \\ &= \frac{1}{3} \begin{pmatrix} 32.0 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix} \end{aligned}$$

2.4 Normal Distribution

The normal distribution is one of the most important probability density functions, especially since many physically observed variables follow an approximately normal distribution. Furthermore, the sampling distribution of the mean of any arbitrary probability distribution follows a normal distribution. The normal distribution also plays an important role as the parametric distribution of choice in clustering, density estimation, and classification.

2.4.1 Univariate Normal Distribution

A random variable X has a normal distribution, with the parameters mean μ and variance σ^2, if the probability density function of X is given as follows

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2.52}$$

The term $(x-\mu)^2$ measures the distance of a value x from the mean μ of the distribution, and thus the probability density decreases exponentially as a function of the distance from the mean. The maximum value of the density occurs at the mean value $x = \mu$, given as $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}$, which is inversely proportional to the standard deviation σ of the distribution.

Figure 2.6 plots the standard normal distribution, which has the parameters $\mu = 0$ and $\sigma^2 = 1$. The normal distribution has a characteristic *bell* shape, and it is symmetric about the mean. The figure also shows the effect of different values of standard deviation on the shape of the distribution. A smaller value (e.g., $\sigma = 0.5$) results in a more "peaked" distribution that decays faster, whereas a larger value (e.g., $\sigma = 2$) results in a flatter distribution that decays slower. Since the normal distribution is symmetric, the mean μ is also the median, as well as the mode, of the distribution.

Probability Mass Given an interval $[a, b]$ the probability mass of the Normal distribution within that interval is given as

$$P(a \le x \le b) = \int_a^b f(x|\ \mu, \sigma^2)dx$$

In particular, we are often interested in the probability mass concentrated within k standard deviations from the mean, i.e., for the interval $[\mu - k\sigma, \mu + k\sigma]$, which can be computed as

$$P(\mu - k\sigma \le x \le \mu + k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-k\sigma}^{\mu+k\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

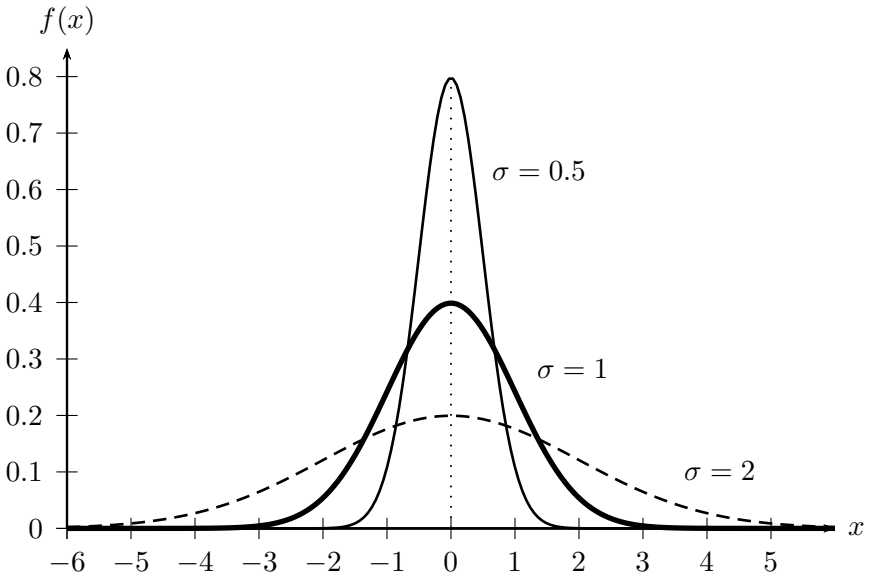

Figure 2.6: Normal Distribution: $\mu = 0$, and different variances

Via a change of variable $z = \frac{x-\mu}{\sigma}$, we get an equivalent formulation in terms of the standard normal distribution

$$P(-k \leq z \leq k) = \frac{1}{\sqrt{2\pi}} \int_{-k}^{k} e^{-\frac{1}{2}z^2} dz$$
$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{k} e^{-\frac{1}{2}z^2} dz$$

The last step follows from the fact that $e^{-\frac{1}{2}z^2}$ is symmetric, and thus the integral over the range $[-k, k]$ is equivalent to 2 times the integral over the range $[0, k]$. Finally, via another change of variable $t = \frac{z}{\sqrt{2}}$, we get

$$P(-k \leq z \leq k) = P(0 \leq t \leq k/\sqrt{2}) = \frac{2}{\sqrt{\pi}} \int_{0}^{k/\sqrt{2}} e^{-t^2} dt = \text{erf}\left(k/\sqrt{2}\right) \qquad (2.53)$$

where erf is the *Gauss error function*, defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$

Using (2.53) we can compute the probability mass within k standard deviations of the mean. In particular, for $k = 1$, we have $P(\mu - \sigma \leq x \leq \mu + \sigma) = \text{erf}(1/\sqrt{2}) =$

0.6827, which means that 68.27% of all points lie within one standard deviation from the mean. For $k = 2$, we have $\text{erf}(2/\sqrt{2}) = 0.9545$, and for $k = 3$ we have $\text{erf}(3/\sqrt{2}) = 0.9973$. Thus almost the entire probability mass (99.73%) of a normal distribution is within $\pm 3\sigma$ from the mean μ.

2.4.2 Multivariate Normal Distribution

Given the d-dimensional vector random variable $\mathbf{X} = (X_1, X_2, \cdots, X_d)^T$, we say that $\mathbf{X}$ has a multivariate normal distribution, with the parameters mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, if its joint multivariate probability density function is given as follows

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right\} \tag{2.54}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix. As in the univariate case, the term

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \tag{2.55}$$

measures the distance, called the *Mahalanobis distance*, of the point $\mathbf{x}$ from the mean $\boldsymbol{\mu}$ of the distribution, taking into account all of the variance-covariance information between the attributes. The Mahalanobis distance is a generalization of Euclidean distance, since if we set $\boldsymbol{\Sigma} = \mathbf{I}$, where $\mathbf{I}$ is the $d \times d$ identity matrix (with diagonal elements as 1's and off-diagonal elements as 0's), we get

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{I}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$$

The Euclidean distance thus ignores the covariance information between the attributes, whereas the Mahalanobis distance explicitly takes it into consideration.

The standard multivariate normal distribution has parameters $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$. Figure 2.7(a) plots the probability density of the standard bivariate ($d = 2$) normal distribution, with parameters

$$\boldsymbol{\mu} = \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This corresponds to the case where the two attributes are independent, and both follow the standard normal distribution. The symmetric nature of the standard normal distribution can be clearly seen in the contour plot shown in Figure 2.7(b). Each level curve represents the set of points $\mathbf{x}$ with a fixed density value $f(\mathbf{x})$.

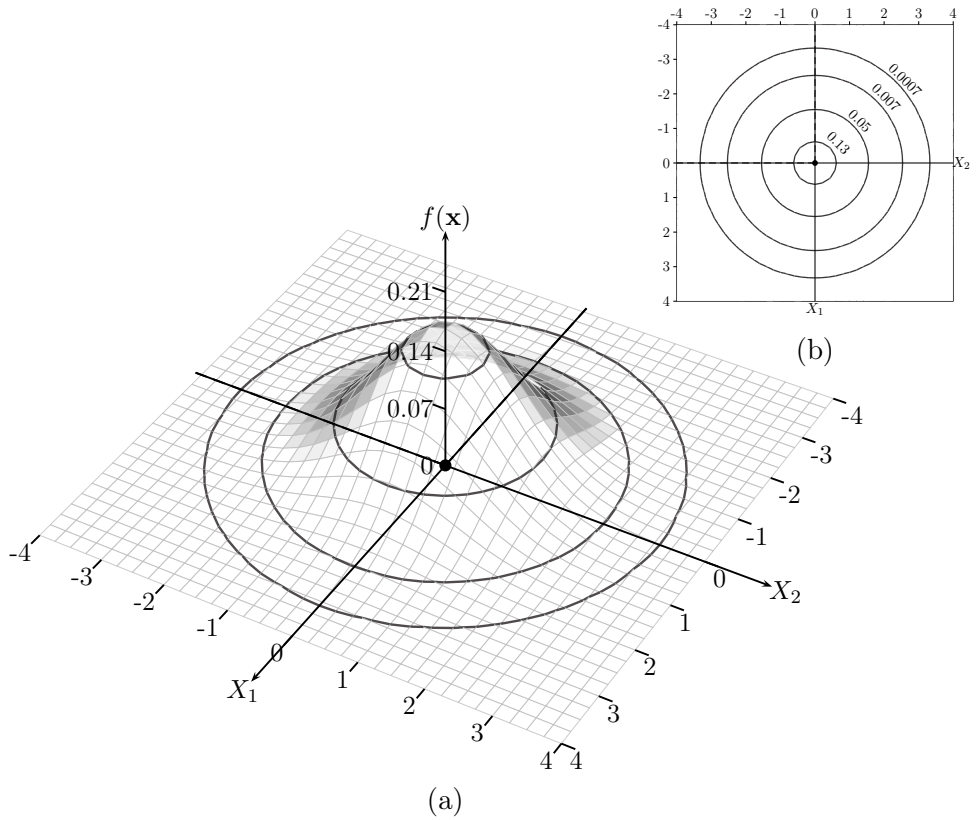

Figure 2.7: Standard Bivariate Normal Density (a) and its Contour Plot (b). Parameters: $\boldsymbol{\mu} = (0, 0)^T$, $\boldsymbol{\Sigma} = \mathbf{I}$

Geometry of the Multivariate Normal Let us consider the geometry of the multivariate Normal distribution for an arbitrary given mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Compared to the standard Normal distribution, we can expect the density contours to be shifted, scaled and rotated. The shift or translation comes from the fact that the mean $\boldsymbol{\mu}$ is not necessarily the origin $\mathbf{0}$. The scaling or skewing is a result of the attribute variances, and the rotation is a result of the covariances.

The shape or geometry of the normal distribution becomes clear by considering the eigen-decomposition of the covariance matrix. Recall that $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric positive semidefinite matrix. The eigenvector equation for $\boldsymbol{\Sigma}$ is given as

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{2.56}$$

Here λ_i is an eigenvalue of $\boldsymbol{\Sigma}$ and the vector $\mathbf{u}_i \in \mathbb{R}^d$ is the eigenvector corresponding to λ_i. Since $\boldsymbol{\Sigma}$ is symmetric and positive semidefinite it has d real and non-negative eigenvalues, which can be arranged in order from the largest to the smallest as follows:

$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$. The diagonal matrix $\mathbf{\Lambda}$ is used to record these eigenvalues

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

Further, the eigenvectors are unit vectors (normal) and are mutually orthogonal, i.e.,

$$\mathbf{u}_i^T \mathbf{u}_i = 1 \quad \text{for all } i$$
$$\mathbf{u}_i^T \mathbf{u}_j = 0 \quad \text{for all } i \neq j$$

The eigenvectors can be put together into an orthonormal matrix $\mathbf{U}$, defined as a matrix with normal and mutually orthogonal columns

$$\mathbf{U} = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & & | \end{pmatrix}$$

The eigen-decomposition of $\mathbf{\Sigma}$ can then be expressed compactly as follows

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{2.57}$$

This equation can be interpreted geometrically as a change in basis vectors. From the original d dimensions corresponding to the d attributes X_j, we derive d new dimensions $\mathbf{u}_i$. $\mathbf{\Sigma}$ is the covariance matrix in the original space, whereas $\mathbf{\Lambda}$ is the covariance matrix in the new coordinate space. Since $\mathbf{\Lambda}$ is a diagonal matrix, we can immediately conclude that after the transformation, each new dimension $\mathbf{u}_i$ has a variance λ_i, and further that all covariances are zero. In other words, in the new space, the normal distribution is axis aligned (has no rotation component), but is skewed in each axis proportional to the eigenvalue λ_i, which represents the variance along that dimension.

Total and Generalized Variance: The determinant of the covariance matrix is is given as $det(\mathbf{\Sigma}) = \prod_{i=1}^{d} \lambda_i$. Thus, the generalized variance of $\mathbf{\Sigma}$ is the product of its eigenvectors.

Given the fact that the trace of square matrix is invariant to similarity transformation, such as a change of basis, we conclude that the total variance $var(\mathbf{D})$ for a dataset $\mathbf{D}$ is invariant, i.e.,

$$var(\mathbf{D}) = tr(\mathbf{\Sigma}) = \sum_{i=1}^{d} \sigma_i^2 = \sum_{i=1}^{d} \lambda_i = tr(\Lambda)$$

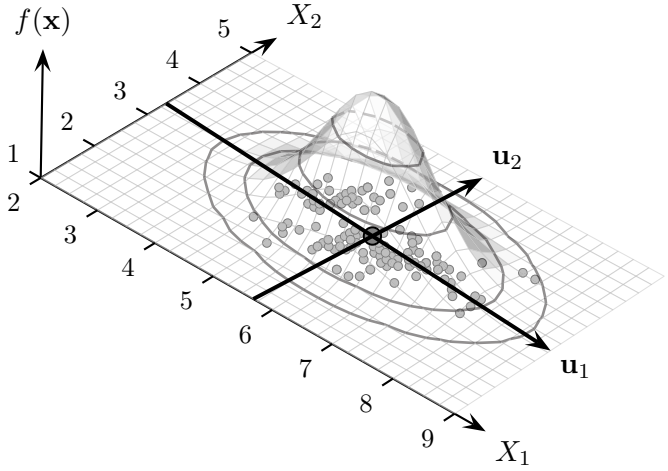

Figure 2.8: Iris: `sepal length` and `sepal width`, Bivariate Normal Density and Contours

Example 2.6 (Bivariate Normal Density): Treating attributes `sepal length` (X_1) and `sepal width` (X_2) in the Iris dataset (see Table 1.1) as continuous random variables, we can define a continuous bivariate random variable $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Assuming that $\mathbf{X}$ follows a bivariate normal distribution, we can estimate its parameters from the sample. The sample mean is given as

$$\hat{\boldsymbol{\mu}} = (5.843, 3.054)^T$$

and the sample covariance matrix is given as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The plot of the bivariate normal density for the two attributes is shown in Figure 2.8. The figure also shows the contour lines and the data points.

Consider the point $\mathbf{x}_2 = (6.9, 3.1)^T$. We have

$$\mathbf{x}_2 - \hat{\boldsymbol{\mu}} = \begin{pmatrix} 6.9 \\ 3.1 \end{pmatrix} - \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix} = \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix}$$

The Mahalanobis distance between $\mathbf{x}_2$ and $\hat{\boldsymbol{\mu}}$ is

$$\begin{aligned} (\mathbf{x}_i - \boldsymbol{\mu})^T \, \widehat{\boldsymbol{\Sigma}}^{-1} \, (\mathbf{x}_i - \boldsymbol{\mu}) &= \begin{pmatrix} 1.057 & 0.046 \end{pmatrix} \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}^{-1} \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} \\ &= \begin{pmatrix} 1.057 & 0.046 \end{pmatrix} \begin{pmatrix} 1.486 & 0.31 \\ 0.31 & 5.42 \end{pmatrix} \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} \\ &= 1.701 \end{aligned}$$

whereas the squared Euclidean distance between them is

$$\|(\mathbf{x}_2 - \hat{\boldsymbol{\mu}})\|^2 = \begin{pmatrix} 1.057 & 0.046 \end{pmatrix} \begin{pmatrix} 1.057 \\ 0.046 \end{pmatrix} = 1.119$$

The eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ are given as: $\lambda_1 = 0.684$, and $\lambda_2 = 0.184$, with the corresponding eigenvectors

$$\mathbf{u}_1 = (-0.997, 0.078)^T$$
$$\mathbf{u}_2 = (-0.078, -0.997)^T$$

These two eigenvectors define the new axes in which the covariance matrix is given as

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.684 & 0 \\ 0 & 0.184 \end{pmatrix}$$

The angle between the original axes $\mathbf{e}_1 = (1, 0)^T$ and $\mathbf{u}_1$ specifies the rotation angle for the multivariate normal

$$\cos\theta = \mathbf{e}_1^T \mathbf{u}_1 = -0.997$$
$$\theta = cos^{-1}(-0.997) = 175.5^\circ$$

Figure 2.8 illustrates the new coordinate axes and the new variances. We can see that in the original axes, the contours are only slightly rotated (by angle 4.5°).

2.5 Annotated References

2.6 Exercises and Projects

1. Answer the following:
   (a) What is the difference between a Model and a Pattern?
   (b) What is the difference between Description and Prediction?
   (c) What are the four components of a data mining algorithm?
   (d) True or False:
      i. Mean is robust against outliers.
      ii. Median is robust against outliers.
      iii. Standard Deviation is robust against outliers.