Sample Mean Vector and Sample Covariance Matrix

1 Sample mean and sample covariance

Recall that in 1-dimensional case, in a sample $x_1, \dots, x_n$, we can define

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

as the (unbiased) sample mean

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

p-dimensional case: Suppose we have p variates $X_1, \dots, X_p$. For the vector of variates

$$\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix},$$

we have a p-variate sample with size n:

$$\vec{x}_1, \dots, \vec{x}_n \in \mathbb{R}^p.$$

This sample of n observations give the following data matrix:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix}. \tag{1.1}$$

Notice that here each column in the data matrix corresponds to a particular variate X_j.

Sample mean: For each variate X_j, define the sample mean:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \ j = 1, \dots, p.$$

Then the sample mean vector

$$\bar{\vec{x}} := \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum\limits_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n} \sum\limits_{i=1}^{n} x_{ip} \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i.$$

This can be further represented as

$$\bar{\vec{x}} = \frac{1}{n} [\vec{x}_1, \dots, \vec{x}_n] \vec{1}_n = \frac{1}{n} \boldsymbol{X}^\top \vec{1}_n.$$

Sample covariance matrix: For each variate X_j, $j = 1, \dots, p$, define its sample variance as

$$s_{jj} = s_j^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2, \ j = 1, \dots, p$$

and sample covariance between X_j and X_k

$$s_{jk} = s_{kj} := \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), 1 \le k, j \le p, \ j \ne k.$$

The sample covariance matrix is defined as

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix},$$

Then

$$\begin{aligned} \boldsymbol{S} &= \begin{bmatrix} \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 & \dots & \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ \frac{1}{n-1} \sum_{i=1}^{n} (x_{ip} - \bar{x}_p)(x_{i1} - \bar{x}_1) & \dots & \frac{1}{n-1} \sum_{i=1}^{n} (x_{ip} - \bar{x}_p)^2 \end{bmatrix} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \begin{bmatrix} (x_{i1} - \bar{x}_1)^2 & \dots & (x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{ip} - \bar{x}_p)(x_{i1} - \bar{x}_1) & \dots & (x_{ip} - \bar{x}_p)^2 \end{bmatrix} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \begin{bmatrix} x_{i1} - \bar{x}_1 \\ \vdots \\ x_{ip} - \bar{x}_p \end{bmatrix} \begin{bmatrix} x_{i1} - \bar{x}_1 & \dots & x_{ip} - \bar{x}_p \end{bmatrix} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i - \bar{\vec{x}} \right) \left(\vec{x}_i - \bar{\vec{x}} \right)^\top. \end{aligned}$$

Define the centered data matrix as

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{bmatrix} = \boldsymbol{X} - \vec{1}_n \bar{\vec{x}}^\top$$

We can also represent $\boldsymbol{S}$ as

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{n-1} [\vec{x}_1 - \bar{\vec{x}}, \dots, \vec{x}_n - \bar{\vec{x}}] \begin{bmatrix} \vec{x}_1^\top - \bar{\vec{x}}^\top \\ \vdots \\ \vec{x}_n^\top - \bar{\vec{x}}^\top \end{bmatrix} \\ &= \frac{1}{n-1} (\boldsymbol{X} - \vec{1}_n \bar{\vec{x}}^\top)^\top (\boldsymbol{X} - \vec{1}_n \bar{\vec{x}}^\top) \end{aligned}$$

2 Linear transformation of observations

Consider a sample of $\vec{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ with size n:

$$\vec{x}_1, \dots, \vec{x}_n.$$

The corresponding data matrix is represented as

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix}.$$

For some $\boldsymbol{C} \in \mathbb{R}^{q\times p}$ and $\vec{d} \in \mathbb{R}^q$, consider the linear transformation

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix} = \boldsymbol{C}\vec{X} + \vec{d}.$$

Then we get a q-variate sample:

$$\vec{y}_i = \boldsymbol{C}\vec{x}_i + \vec{d}, \quad i = 1, \dots, n,$$

The sample mean of $\vec{y}_1, \dots, \vec{y}_n$ is

$$\overline{\vec{y}} = \frac{1}{n}\sum_{i=1}^{n} \vec{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{C}\vec{x}_i + \vec{d}) = \boldsymbol{C}(\frac{1}{n}\sum_{i=1}^{n}\vec{x}_i) + \vec{d} = \boldsymbol{C}\overline{\vec{x}} + \vec{d}.$$

And the sample covariance is

$$\begin{aligned} \boldsymbol{S}_y &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{y}_i - \overline{\vec{y}}\right)\left(\vec{y}_i - \overline{\vec{y}}\right)^\top \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{C}\vec{x}_i - \boldsymbol{C}\overline{\vec{x}}\right)\left(\boldsymbol{C}\vec{x}_i - \boldsymbol{C}\overline{\vec{x}}\right)^\top \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{C}\left(\vec{x}_i - \overline{\vec{x}}\right)\left(\vec{x}_i - \overline{\vec{x}}\right)^\top \boldsymbol{C}^\top \\ &= \boldsymbol{C}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\vec{x}_i - \overline{\vec{x}}\right)\left(\vec{x}_i - \overline{\vec{x}}\right)^\top\right)\boldsymbol{C}^\top \\ &= \boldsymbol{C}\boldsymbol{S}_x\boldsymbol{C}^\top. \end{aligned}$$

3 Block structure of the sample covariance

For the vector $\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$, we can divide it into two parts: $\vec{X}^{(1)} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix}$ and $\vec{X}^{(2)} = \begin{bmatrix} X_{q+1} \\ X_{q+2} \\ \vdots \\ X_p \end{bmatrix}$. In other words,

$$\vec{X} = \left[\begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_q \\ \hdashline X_{q+1} \\ X_{q+2} \\ \vdots \\ X_p \end{array}\right] = \left[\begin{array}{c} \vec{X}^{(1)} \\ \hdashline \vec{X}^{(2)} \end{array}\right]$$

For a sample $\vec{x}_1, \ldots, \vec{x}_n$ of $\vec{X}$, we have the partition

$$\vec{x}_i = \left[\begin{array}{c} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iq} \\ \hdashline x_{i(q+1)} \\ x_{i(q+2)} \\ \vdots \\ x_{ip} \end{array}\right] = \left[\begin{array}{c} \vec{x}_i^{(1)} \\ \hdashline \vec{x}_i^{(2)} \end{array}\right]$$

where

$$\vec{x}_i^{(1)} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iq} \end{bmatrix}, \quad \vec{x}_i^{(2)} = \begin{bmatrix} x_{i(q+1)} \\ x_{i(q+2)} \\ \vdots \\ x_{ip} \end{bmatrix}$$

We have the partition of the sample mean directly

$$\bar{\vec{x}} = \left[\begin{array}{c} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_q \\ \hdashline \bar{x}_{q+1} \\ \bar{x}_{q+2} \\ \vdots \\ \bar{x}_p \end{array}\right] = \left[\begin{array}{c} \bar{\vec{x}}^{(1)} \\ \hdashline \bar{\vec{x}}^{(2)} \end{array}\right].$$

Furthermore, we partition the sample covariance in the following manner:

$$\boldsymbol{S} = \left[\begin{array}{ccc:ccc} s_{11} & \cdots & s_{1q} & s_{1,q+1} & \cdots & s_{1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_{q1} & \cdots & s_{qq} & s_{q,q+1} & \cdots & s_{q,p} \\ \hdashline s_{q+1,1} & \cdots & s_{q+1,q} & s_{q+1,q+1} & \cdots & s_{q+1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pq} & s_{p,q+1} & \cdots & s_{p,p} \end{array}\right] = \left[\begin{array}{c:c} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \hdashline \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{array}\right].$$

By definition, $\boldsymbol{S}_{11}$ is the sample covariance of $\vec{X}^{(1)}$ and $\boldsymbol{S}_{22}$ is the sample covariance of $\vec{X}^{(2)}$. Here $\boldsymbol{S}_{12}$ is referred to as the sample cross covariance matrix between $\vec{X}^{(1)}$ and $\vec{X}^{(2)}$. In fact, we can derive the following formula:

$$\boldsymbol{S}_{21} = \boldsymbol{S}_{12}^{\top} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{x}_i^{(2)} - \bar{\vec{x}}^{(2)}\right) \left(\vec{x}_i^{(1)} - \bar{\vec{x}}^{(1)}\right)^{\top}$$

4 Standardization and Sample Correlation Matrix

For the data matrix (1.1). The sample mean vector is denoted as $\bar{\vec{x}}$ and the sample covariance is denoted as $\boldsymbol{S}$. In particular, for $j = 1, \ldots, p$, let $\bar{x}_j$ be the sample mean of the j-th variable and $\sqrt{s_{jj}}$ be the sample standard deviation.

For any entry x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, we get the standardized entry

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}}.$$

Then the data matrix $\boldsymbol{X}$ is standardized to

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} \vec{z}_1^\top \\ \vec{z}_2^\top \\ \vdots \\ \vec{z}_n^\top \end{bmatrix}.$$

Denote by $\boldsymbol{R}$ the sample covariance for the sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. What is the connection between $\boldsymbol{R}$ and $\boldsymbol{S}$?

The i-th row of $\boldsymbol{Z}$ can be written as

$$\begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ip} \end{bmatrix} = \begin{bmatrix} (x_{i1} - \bar{x}_1)/\sqrt{s_{11}} \\ (x_{i2} - \bar{x}_2)/\sqrt{s_{22}} \\ \vdots \\ (x_{ip} - \bar{x}_p)/\sqrt{s_{pp}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & & & \\ & \frac{1}{\sqrt{s_{22}}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix} \begin{bmatrix} x_{i1} - \bar{x}_1 \\ x_{i2} - \bar{x}_2 \\ \vdots \\ x_{ip} - \bar{x}_p \end{bmatrix}$$

Let

$$\boldsymbol{V}^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & & & \\ & \frac{1}{\sqrt{s_{22}}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}.$$

This transformation can be represented as

$$\vec{z}_i = \boldsymbol{V}^{-\frac{1}{2}}(\vec{x}_i - \bar{\vec{x}}) = \boldsymbol{V}^{-\frac{1}{2}}\vec{x}_i - \boldsymbol{V}^{-\frac{1}{2}}\bar{\vec{x}}, \quad i = 1, \ldots, n.$$

This implies that the sample mean for the new data matrix is

$$\bar{\vec{z}} = \boldsymbol{V}^{-\frac{1}{2}}(\bar{\vec{x}} - \bar{\vec{x}}) = \vec{0},$$

By the formula for the sample covariance of linear combinations of variates, the sample covariance matrix for the new data matrix $\boldsymbol{Z}$ is

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{S}\left(\boldsymbol{V}^{-\frac{1}{2}}\right)^\top \\ &= \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & & & \\ & \frac{1}{\sqrt{s_{22}}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & & & \\ & \frac{1}{\sqrt{s_{22}}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{s_{12}}{\sqrt{s_{11}s_{22}}} & \cdots & \frac{s_{1p}}{\sqrt{s_{11}s_{pp}}} \\ \frac{s_{21}}{\sqrt{s_{22}s_{11}}} & 1 & \cdots & \frac{s_{2p}}{\sqrt{s_{22}s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{s_{p1}}{\sqrt{s_{pp}s_{11}}} & \frac{s_{p2}}{\sqrt{s_{pp}s_{22}}} & \cdots & 1 \end{bmatrix} := \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \end{aligned}$$

The matrix $\boldsymbol{R}$ is called the sample correlation matrix for the original data matrix $\boldsymbol{X}$.

5 Mahalanobis distance and mean-centered ellipse

Sample covariance is p.s.d.

Recall that the sample covariance is

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\vec{x}_i - \bar{\vec{x}})(\vec{x}_i - \bar{\vec{x}})^\top.$$

Is $\boldsymbol{S}$ always positive semidefinite? Consider the spectral decomposition

$$\boldsymbol{S} = \sum_{j=1}^{p}\lambda_j \vec{u}_j \vec{u}_j^\top.$$

Then $\boldsymbol{S}\vec{u}_j = \lambda_j \vec{u}_j$, which implies that

$$\vec{u}_j^\top \boldsymbol{S}\vec{u}_j = \vec{u}_j^\top(\lambda_j \vec{u}_j) = \lambda_j \vec{u}_j^\top \vec{u}_j = \lambda_j.$$

On the other hand

$$\begin{aligned}
\vec{u}_j^\top \boldsymbol{S}\vec{u}_j &= \frac{1}{n-1}\vec{u}_j^\top\left(\sum_{i=1}^{n}(\vec{x}_i - \bar{\vec{x}})(\vec{x}_i - \bar{\vec{x}})^\top\right)\vec{u}_j \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\vec{u}_j^\top(\vec{x}_i - \bar{\vec{x}})(\vec{x}_i - \bar{\vec{x}})^\top \vec{u}_j \\
&= \frac{1}{n-1}\sum_{i=1}^{n}|\vec{u}_j^\top(\vec{x}_i - \bar{\vec{x}})|^2 \geq 0.
\end{aligned}$$

This implies that all eigenvalues of $\boldsymbol{S}$ are nonnegative, so $\boldsymbol{S}$ is positive semidefinite.

In this course, we always assume $n > p$ and $\boldsymbol{S}$ is positive definite, which also implies that the inverse sample covariance matrix $\boldsymbol{S}^{-1}$ is also positive definite.

Mahalanobis distance

For any two vectors $\vec{x}, \vec{y} \in \mathbb{R}^p$, their Mahalanobis distance based on $\boldsymbol{S}^{-1}$ is defined as

$$d_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^\top \boldsymbol{S}^{-1}(\vec{x} - \vec{y})}.$$

By spectral decomposition of $\boldsymbol{S}^{-1}$:

$$\boldsymbol{S}^{-1} = \sum_{j=1}^{p}\frac{1}{\lambda_j}\vec{u}_j \vec{u}_j^\top,$$

the Mahalanobis distance is well-defined since

$$(\vec{x} - \vec{y})^\top \boldsymbol{S}^{-1}(\vec{x} - \vec{y}) = (\vec{x} - \vec{y})^\top\left(\sum_{j=1}^{p}\frac{1}{\lambda_j}\vec{u}_j \vec{u}_j^\top\right)(\vec{x} - \vec{y}) = \sum_{j=1}^{p}\frac{1}{\lambda_j}|(\vec{x} - \vec{y})^\top \vec{u}_j|^2 \geq 0.$$

The mean-centered ellipse with Mahalanobis radius c is defined as

$$\{\vec{x} \in \mathbb{R}^p : d_M(\vec{x}, \bar{\vec{x}}) \leq c\} = \{\vec{x} \in \mathbb{R}^p : (\vec{x} - \bar{\vec{x}})^\top \boldsymbol{S}^{-1}(\vec{x} - \bar{\vec{x}}) \leq c^2\}.$$

Mean-centered ellipse

For any $\vec{x}$, we have

$$(\vec{x}-\bar{\vec{x}})^\top \boldsymbol{S}^{-1}(\vec{x}-\bar{\vec{x}}) = (\vec{x}-\bar{\vec{x}})^\top \left(\sum_{j=1}^{p} \frac{1}{\lambda_j}\vec{u}_j\vec{u}_j^\top\right)(\vec{x}-\bar{\vec{x}}) = \sum_{j=1}^{p}\frac{1}{\lambda_j}|(\vec{x}-\bar{\vec{x}})^\top\vec{u}_j|^2$$

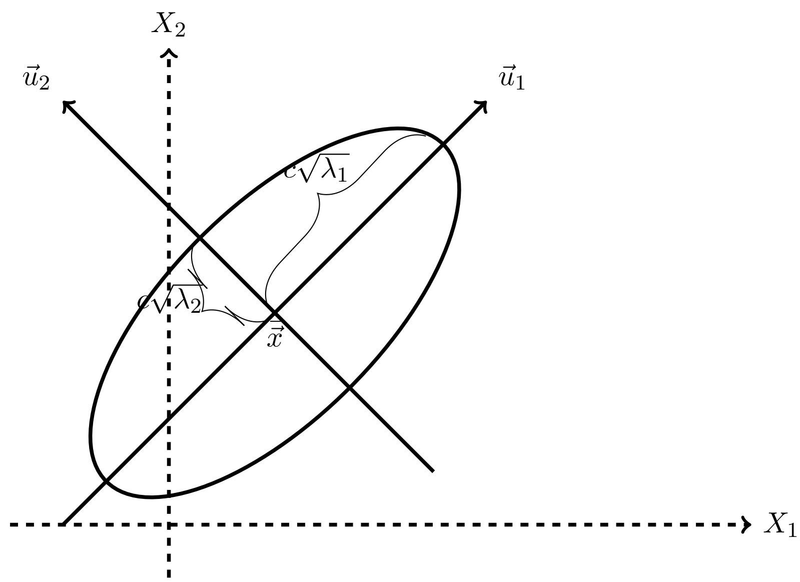
Consider a new cartesian coordinate system with center $\bar{\vec{x}}$ and axes $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_p$, the new coordinates of $\vec{x}$ based on the axis $\vec{u}_j$ becomes $w_j = (\vec{x}-\bar{\vec{x}})^\top\vec{u}_j$, j=1, $\ldots$, p. Then the mean-centered ellipse

$$\{\vec{x} : (\vec{x}-\bar{\vec{x}})^\top \boldsymbol{S}^{-1}(\vec{x}-\bar{\vec{x}}) \le c^2\}$$

becomes

$$\{\vec{w} : \sum_{j=1}^{p}\frac{1}{(\sqrt{\lambda_j})^2}w_j^2 \le c^2\} = \{\vec{w} : \sum_{j=1}^{p}\frac{1}{(c\sqrt{\lambda_j})^2}w_j^2 \le 1\}$$

in the new coordinate system, which is an ellipse with half axis lengths $c\sqrt{\lambda_1}$, $c\sqrt{\lambda_2}$, ..., $c\sqrt{\lambda_p}$.



6 Examples

Example 1

Consider a 2-variate data matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}$$

with sample mean vector $\overline{\vec{x}}$ and sample covariance matrix $\boldsymbol{S}_{\vec{x}}$.

Define the new sample

$$y_1 = x_{11} + x_{12}, y_2 = x_{21} + x_{22}, ..., y_n = x_{n1} + x_{n2}.$$

Can we compute its sample mean and sample variance directly through $\overline{\vec{x}}$ and $\boldsymbol{S}_{\vec{x}}$?

Denote $\boldsymbol{C} = [1, 1]$. Then

$$y_i = x_{i1} + x_{i2} = [1, 1] \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} = \boldsymbol{C}\vec{x}_i.$$

The sample mean of $y_1, \ldots, y_n$ can be represented as

$$\begin{aligned} \bar{y} &= \frac{1}{n}\left[(x_{11} + x_{12}) + \ldots + (x_{n1} + x_{n2})\right] \\ &= \frac{1}{n}\left[x_{11} + \ldots + x_{n1}\right] + \frac{1}{n}\left[x_{12} + \ldots + x_{n2}\right] \\ &= \overline{x}_1 + \overline{x}_2 \\ &= \boldsymbol{C}\overline{\vec{x}}. \end{aligned}$$

Represent the sample variance of $y_1, \ldots, y_n$ by s_y^2. Then

$$\begin{aligned} (n-1)s_y^2 &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}\left((x_{i1} + x_{i2}) - (\overline{x}_1 + \overline{x}_2)\right)^2 \\ &= \sum_{i=1}^{n}\left((x_{i1} - \overline{x}_1) + (x_{i2} - \overline{x}_2)\right)^2 \\ &= \sum_{i=1}^{n}\left((x_{i1} - \overline{x}_1)^2 + 2(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) + (x_{i2} - \overline{x}_2)^2\right) \\ &= \sum_{i=1}^{n}(x_{i1} - \overline{x}_1)^2 + 2\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) + \sum_{i=1}^{n}(x_{i2} - \overline{x}_2)^2 \\ &= (n-1)s_{11} + 2(n-1)s_{12} + (n-1)s_{22}. \end{aligned}$$

Then

$$\begin{aligned} s_y^2 &= s_{11} + 2s_{12} + s_{22} = s_{11} + s_{12} + s_{21} + s_{22} \\ &= [1, 1]\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^\top \end{aligned}$$

Example 2

Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times 4}$ is a data matrix for the variables $\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$, with the following sample covariance

$$\boldsymbol{S}_x = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

What is the sample cross-covariance matrix between $\begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ and $\begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$?

Solution Since

$$\vec{Y} := \begin{bmatrix} X_1 \\ X_3 \\ X_2 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} := \boldsymbol{C}\vec{X},$$

we know it sample covariance matrix is

$$\begin{aligned} \boldsymbol{S}_y &= \boldsymbol{C}\boldsymbol{S}_x\boldsymbol{C}^\top \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{bmatrix}. \end{aligned}$$

From the partition

$$\vec{Y} = \left[\begin{array}{c} X_1 \\ X_3 \\ \hdashline X_2 \\ X_4 \end{array}\right]$$

we have the partition

$$\boldsymbol{S}_y = \left[\begin{array}{cc:cc} 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ \hdashline 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{array}\right].$$

Then sample cross-covariance matrix between $\begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ and $\begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$ is $\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$. This result can be verified entrywise.