MATH5745 Multivariate Methods
Lecture 04

Mean, variance, covariance and correlation

February 25, 2019

- In MATH5471 you have discussed the concept of *random variables*
- We only consider *continuous* random variables
- You also have discussed the concept of *probability density function* (pdf), denoted $f(y)$ say.
- The usual convention is capital letter (e.g. X, Y) for random variable, and lower case letter (e.g. x, y) for observed data.
- We use lower case letters for scalar or vector (bold), and capital letter for a matrix.
- We do not distinguish between a random variable and its observed value. (The context will make it clear.)

Mean and variance (Univariate case)

- You have discussed the concept of *expectation* and *variance* of a random variable
- We assume that the univariate population (from which the data are sampled or observed) can be characterised by two parameters: μ (mean) and σ^2 (variance)
- Suppose $y_1, y_2, \ldots, y_n$ is a random *sample*.
- We have that $\mathrm{E}[y_i] = \mu$ for each i (as the corresponding random variable!).
- The sample mean is defined as

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \left(y_1 + y_2 + \ldots + y_n \right).$$

- We have $\mathrm{E}[\overline{y}] = \mu$ (as the corresponding random variable!).

Mean and variance (univariate)

- The variance of the population is defined as

$$\mathrm{Var}[y] = \sigma^2 = \mathrm{E}[(y-\mu)^2] = \mathrm{E}[y^2] - \mu^2.$$

- $\mathrm{Var}[y]$ is the average squared deviation from the mean μ.
- The *sample* variance is calculated as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

- In practice, we commonly use the equivalent formula

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right).$$

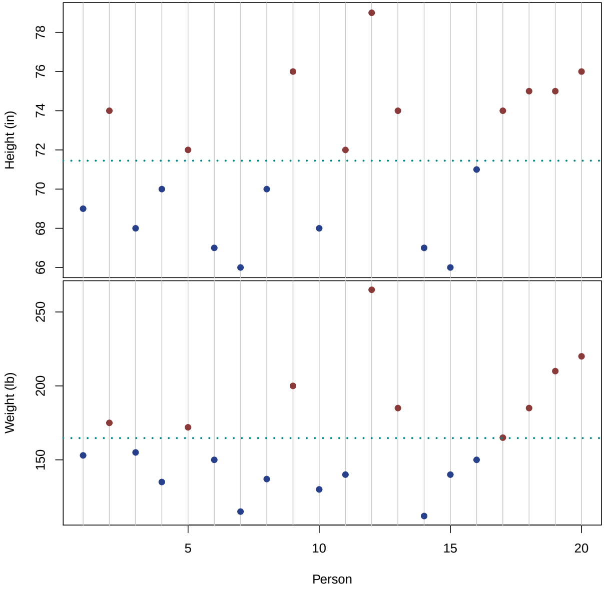
- The *sample* standard deviation is the square root of the sample variance, $s = +\sqrt{s^2}$.
- It can be shown that $\mathrm{E}[s^2] = \sigma^2$ (as the corresponding random variable!).

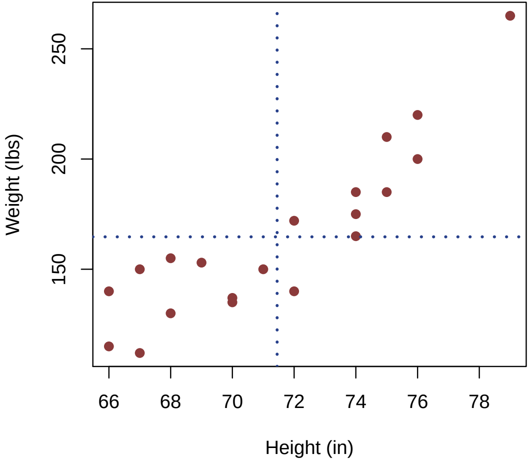
- Let $z_i = ay_i$ for <u>*all*</u> i, where a is a constant.
- Then $\mathrm{E}[z_i] = a\mu_y$ and $\mathrm{Var}[z_i] = a^2\sigma_y^2$ (as random variables!).
- If $\overline{z}$ is the sample mean of z_i, then $\overline{z} = a\overline{y}$ and $\mathrm{E}[\overline{z}] = a\mu_y$.
- If s_z^2 is the sample variance of z_i, then $s_z^2 = a^2 s_y^2$.
- Example: $\mathbf{y} = (3, 7, 8)'$ and $a = 2$ (multiplication constant) so that $\mathbf{z} = (6, 14, 16)'$.
  You can verify that $\overline{y} = 6$, $s_y^2 = 7$, $\overline{z} = 12$, $s_z^2 = 28$.

Covariance and correlation (bivariate)

- Now, consider two variables x and y measured on each subject.
- We have here *bivariate* random variables.
- x and y tend to *vary* together, *covariation*.
- Example: Observe the height (x) and weight (y) of 20 college-age males.

Person	x (height) (inches)	y (weight) lbs.
1	69	153
2	74	175
3	68	155
4	70	135
5	72	172
$\vdots$	$\vdots$	$\vdots$


Height (in)
66
68
70
72
74
76
78
Weight (lb)
150
200
250
5
10
15
20
Person


250
200
150
Weight (lbs)
66
68
70
72
74
76
78
Height (in)

Covariance and correlation (bivariate)

- The *population* covariance between x and y is defined as

$$\mathrm{cov}(x, y) = \sigma_{xy} = \mathrm{E}\left[(x - \mu_x)(y - \mu_y)\right], \text{ or}$$

$$\sigma_{xy} = \mathrm{E}[xy] - \mu_x \mu_y.$$

- Suppose we have n observations in our random sample, the *sample* covariance between x and y is defined as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

or

$$s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \right).$$

Covariance and correlation (bivariate): Example

- Consider the height-weight of the 20 males again.
- Here $\overline{x} = 71.45$ inches and $\overline{y} = 164.7$ lbs.
- We have $\sum_i x_i y_i = 237805$, so that

$$s_{xy} = \frac{1}{19}\{237805 - 20(71.45)(164.7)\} = 128.88.$$

- [QUESTION:] Is this large or small?
- Notice s_{xy} has units "inches lbs." so $s_{xy} = 128.88$ inches lbs. Not easy to interpret!

Covariance and correlation (bivariate)

- Consider the *independence* case.
- If x and y are independent of each other, then $\mathrm{E}[xy] = \mathrm{E}[x]\mathrm{E}[y]$, which means

$$
\begin{aligned}
\sigma_{xy} &= \mathrm{E}[xy] - \mu_x\mu_y \\
&= \mathrm{E}[x]\mathrm{E}[y] - \mu_x\mu_y \\
&= \mu_x\mu_y - \mu_x\mu_y = 0.
\end{aligned}
$$

- If x and y are mutually independent, the covariance is zero.
- Note: if the covariance is zero, it does not necessarily imply that x and y are mutually independent!

- The covariance value depends on units of measurements of x and y.
- In the previous example, if the heights and weights were measured in metres and kilograms (instead of inches and lbs) respectively, the covariance will be lower.
- You can verify that the covariance between x^* (m) and y^* (kg) is 1.485.
- *They convey the* <u>*same*</u> *information/message* about x and y.
- We need a measure of (linear) relationship between two variables that is invariant of scale.
- Standardised by their standard deviation.
- This is called *correlation*.

Covariance and correlation (bivariate)

- The *population* correlation of random variables x and y is given by

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\mathrm{E}[(x-\mu_x)(y-\mu_y)]}{\sqrt{\mathrm{E}[(x-\mu_x)^2]}\sqrt{\mathrm{E}[(y-\mu_y)^2]}}.$$

- The correlation ρ_{xy} ranges from -1 to $+1$.
- If random variables x and y are independent, the correlation is zero (because the covariance σ_{xy} is zero).
- The *sample* correlation is given by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Covariance and correlation (bivariate): Example

- Previously, we have $s_{xy} = 128.88$, $\overline{x} = 71.45$, $\sum_i x_i^2 = 102379$ and we calculate

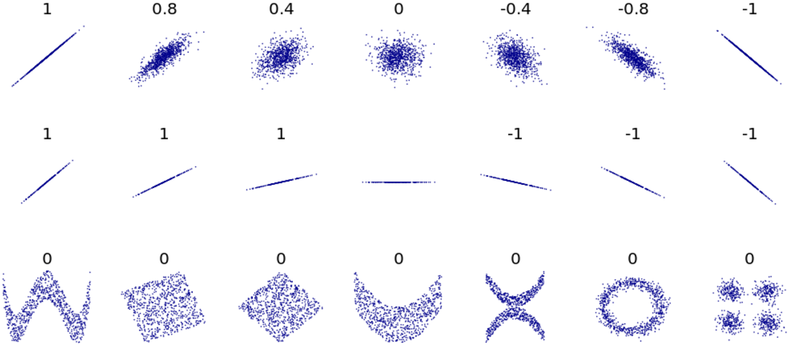
$$s_x^2 = \frac{1}{19}\left(102379 - 20(71.45^2)\right) = 14.576.$$

- Similarly, $s_y^2 = 1441.27$ and then

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{128.88}{\sqrt{14.576}\sqrt{1441.27}} = 0.889.$$

- How meaningful is correlation of 0.889?

Covariance and correlation (bivariate)



Source: Wikipedia

Covariance and correlation (bivariate)

- Recall dot product: For vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}'\mathbf{y} = ||\mathbf{x}||\,||\mathbf{y}||\cos\theta$ where θ is the angle between them.
- The correlation can be interpreted (geometrically) as the cosine of the angle between the two vectors (see the textbook).
- When the correlation is zero, the two vectors are perpendicular (angle$=90^\circ$).

Covariance and correlation (bivariate): Example

- Let $\mathbf{x}' = (-6.64, 15.06, 2.36, -6.94, -3.84)$ and $\mathbf{y}' = (0.92, 2.92, 2.72, 1.12, -7.68)$.
- Note: These have been centred to have zero mean.
- $\mathbf{x}'\mathbf{y} = 66.004$, $\mathbf{x}'\mathbf{x} = 339.372$, $\mathbf{y}'\mathbf{y} = 77.008$.
- $||\mathbf{x}|| = +\sqrt{339.373} = 18.4221$, $||\mathbf{y}|| = \sqrt{77.008} = 8.7754$.
- $\cos\theta = \dfrac{\mathbf{x}'\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} = 0.4083.$
- $s_{xy} = \frac{1}{4}(66.004) = 16.501,$
- $s_x^2 = \frac{1}{4}(339.372) = 84.843, \quad s_x = +\sqrt{s_x^2} = 9.211.$
- $s_y^2 = \frac{1}{4}(77.008) = 19.252, \quad s_y = +\sqrt{s_y^2} = 4.388.$
- $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = 0.4083.$

- Consider the matrix form of data (columns: variables, rows: units of observations):

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}.$$

- Each vector $\mathbf{y}_i$ is a vector of p variables measured on observation i,

$$\mathbf{y}_i' = (y_{i1}, y_{i2}, \ldots, y_{ip}) \quad \text{so} \quad \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix}.$$

- Typically $n > p$.

Sample mean vector

- The sample mean vector $\overline{\mathbf{y}}$ is defined as

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_p \end{pmatrix}$$

  - Notice that:

$$\mathbf{y}_1 + \mathbf{y}_2 + \cdots + \mathbf{y}_n = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1p} \end{pmatrix} + \begin{pmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2p} \end{pmatrix} + \cdots + \begin{pmatrix} y_{n1} \\ y_{n2} \\ \vdots \\ y_{np} \end{pmatrix}.$$

  - The averaging is across units of observation, so $\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$.
  - The vector $\overline{\mathbf{y}}$ is of size $p =$ number of variables.

Sample mean vector

- Using matrix notation, $\bar{\mathbf{y}}$ is given by

$$\bar{\mathbf{y}} = \frac{1}{n}\mathbf{Y}'\mathbf{j}$$

  where $\mathbf{j}$ is an appropriate vector of ones.
- [QUESTION:] What is the length of $\mathbf{j}$ here?
  - Notice that

$$\mathbf{Y}'\mathbf{j} = \begin{pmatrix} y_{11} & y_{21} & \cdots & y_{n1} \\ y_{12} & y_{22} & \cdots & y_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1p} & y_{2p} & \cdots & y_{np} \end{pmatrix} \mathbf{j} = \begin{pmatrix} \sum_i y_{i1} \\ \sum_i y_{i2} \\ \vdots \\ \sum_i y_{ip} \end{pmatrix}.$$

Sample covariance matrix

- The *sample* covariance matrix $\mathbf{S} = (s_{jk})$ is the matrix of the sample variances and covariances of the p variables

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}.$$

- $\mathbf{S}$ is a symmetric (square) $p \times p$ matrix.
- If $\mathbf{S}$ is full rank ($n > p$), $\mathbf{S}$ is positive definite.
  - The sample covariance between variables u and v ($u \neq v$) in $\mathbf{Y}$ is
$$s_{uv} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{iu} - \overline{y}_u)(y_{iv} - \overline{y}_v) = \frac{1}{n-1} \left(\sum_{i=1}^{n} y_{iu} y_{iv} - n \overline{y}_u \overline{y}_v \right).$$
  - s_{uu} is the sample variance of variable u in $\mathbf{Y}$.

Sample mean vector and sample covariance matrix: Example

- Example: For the height-weight data, we have

$$\bar{\mathbf{y}} = \begin{pmatrix} 71.45 \\ 164.7 \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} 14.6 & 128.9 \\ 128.9 & 1441.3 \end{pmatrix}.$$

Sample covariance matrix

- In vector notation:

$$\mathbf{S} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i-\overline{\mathbf{y}})(\mathbf{y}_i-\overline{\mathbf{y}})' \;=\; \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - n\overline{\mathbf{y}}\overline{\mathbf{y}}'\right).$$

  - Notice:
  - $\mathbf{y}_i\mathbf{y}_i' = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix} (y_{i1}, y_{i2}, \ldots, y_{ip}) = \begin{pmatrix} y_{i1}^2 & y_{i1}y_{i2} & \cdots & y_{i1}y_{ip} \\ y_{i2}y_{i1} & y_{i2}^2 & \cdots & y_{i2}y_{ip} \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix}.$
  - Thus $\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' = \begin{pmatrix} \sum y_{i1}^2 & \sum y_{i1}y_{i2} & \cdots & \sum y_{i1}y_{ip} \\ \sum y_{i2}y_{i1} & \sum y_{i2}^2 & \cdots & \sum y_{i2}y_{ip} \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix}.$
  - And similarly can obtain $n\overline{\mathbf{y}}\overline{\mathbf{y}}'$ and so give $\mathbf{S}$ in terms of s_{uv} and s_u^2.

Sample covariance matrix

- In matrix notation:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}.$$

  Here $\mathbf{I}$ is an identity matrix and $\mathbf{J}$ is a matrix of ones.
- [QUESTION:] What is the size of $\mathbf{I}$ and of $\mathbf{J}$?
  - Notice:
  - Write $\mathbf{J} = \mathbf{j}\mathbf{j}'$ so $\mathbf{Y}'\mathbf{J}\mathbf{Y} = (\mathbf{Y}'\mathbf{j})(\mathbf{j}'\mathbf{Y}) = (\mathbf{Y}'\mathbf{j})(\mathbf{Y}'\mathbf{j})' = (n\bar{\mathbf{y}})(n\bar{\mathbf{y}}')$.
  - Also $\mathbf{Y}'\mathbf{I}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)\begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i'$.

Summaries of sample covariance matrix I

- Recall

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}.$$

- Can summarise the "variability" in the data using a univariate quantity.
- Two common summary measures.
  1. Generalised variance, $|\mathbf{S}|$.
     - Same as $\prod_{i=1}^{p} \lambda_i$, where λ_i are eigenvalues of $\mathbf{S}$.
  2. Total variation of $\mathbf{S}$, trace$(\mathbf{S})$.
     - Same as $\sum_{i=1}^{p} \lambda_i$.

Summaries of sample covariance matrix II

- Both summaries are monotonic increasing functions of the eigenvalues.
- They reflect different aspects of the variability in the data.

- $|\mathbf{S}|$ represents the "volume" in $\mathbb{R}^p$ needed to enclose a certain proportion of the data.
- Drawback: if $\lambda_p = 0$, the data is concentrated on a lower dimensional surface and enclosed volume is zero.
- $|\mathbf{S}|$ is useful in maximum likelihood estimation.

- $\text{trace}(\mathbf{S}) = s_{11} + s_{22} + \cdots + s_{pp}$.
- $\text{trace}(\mathbf{S})$ is the sum of variances: "total variation in the data".
- Drawback: $\text{trace}(\mathbf{S})$ ignores covariance (correlation) terms in the data.
- $\text{trace}(\mathbf{S})$ is a useful tool in principal component analysis.

Summaries of sample covariance matrix: Example

- Height-weight data:

$$\mathbf{S} = \begin{pmatrix} 14.6 & 128.9 \\ 128.9 & 1441.3 \end{pmatrix}.$$

- $|\mathbf{S}| = 4427.8$.
- $\text{trace}(\mathbf{S}) = 14.6 + 1441.3 = 1455.9$.

Sample correlation matrix

- The *sample* correlation between variable j and variable k is defined as

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \frac{s_{jk}}{s_j s_k}$$

where $s_j = \sqrt{s_{jj}}$ and $s_k = \sqrt{s_{kk}}$.

- The *sample* correlation matrix is (analogous to the *sample* covariance matrix):

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}.$$

- Correlation between a variable with itself is 1 (!)
- $\mathbf{R}$ is a symmetric (square) $p \times p$ matrix.

Sample correlation matrix

- Example: For the height-weight data, we have

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.889 \\ 0.889 & 1.000 \end{pmatrix}.$$

- Relationship between $\mathbf{S}$ and $\mathbf{R}$:

$$\begin{aligned} \mathbf{R} &= \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1} \\ \mathbf{S} &= \mathbf{D}_s\mathbf{R}\mathbf{D}_s \end{aligned}$$

where $\mathbf{D}_s = \mathrm{diag}(s_1, s_2, \ldots, s_p)$.

Sample correlation matrix: Example

- Example: For the height-weight data, we have

$$\mathbf{S} = \begin{pmatrix} 14.6 & 128.9 \\ 128.9 & 1441.3 \end{pmatrix}.$$

- $\mathbf{D}_s = \mathrm{diag}(\sqrt{14.6}, \sqrt{1441.3}) = \mathrm{diag}(3.821, 37.964)$.
- $\mathbf{D}_s^{-1} = \mathrm{diag}(1/3.821, 1/37.964) = \mathrm{diag}(0.2617, 0.0263)$.
- Here

$$\begin{aligned} \mathbf{R} &= \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1} \\ &= \begin{pmatrix} 0.2617 & 0 \\ 0 & 0.0263 \end{pmatrix} \begin{pmatrix} 14.6 & 128.9 \\ 128.9 & 1441.3 \end{pmatrix} \begin{pmatrix} 0.2617 & 0 \\ 0 & 0.0263 \end{pmatrix} \\ &= \begin{pmatrix} 1.000 & 0.889 \\ 0.889 & 1.000 \end{pmatrix}. \end{aligned}$$

Sample correlation matrix: Example

- Example: For the height-weight data, we have

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.889 \\ 0.889 & 1.000 \end{pmatrix}.$$

- $\mathbf{D}_s = \text{diag}(\sqrt{14.6}, \sqrt{1441.3}) = \text{diag}(3.821, 37.964)$.
- Here

$$\begin{aligned} \mathbf{S} &= \mathbf{D}_s \mathbf{R} \mathbf{D}_s \\ &= \begin{pmatrix} 3.821 & 0 \\ 0 & 37.964 \end{pmatrix} \begin{pmatrix} 1.000 & 0.889 \\ 0.889 & 1.000 \end{pmatrix} \begin{pmatrix} 3.821 & 0 \\ 0 & 37.964 \end{pmatrix} \\ &= \begin{pmatrix} 14.6 & 128.9 \\ 128.9 & 1441.3 \end{pmatrix}. \end{aligned}$$

Sample mean vector and covariance matrix for subset of variables

- Suppose the variables that we have in the data can be naturally grouped into two groups.
- Measured on the same unit of observations (!)
- Example: Several classroom behaviours are observed for students and teachers, with time the unit of observation.
  Aim: study the relationship between pupil and teacher variables.

Sample mean vector and covariance matrix for subset of variables

- Suppose we have p variables (y) and q variables (x).
- Let $\mathbf{y}_i$ be a vector length p and $\mathbf{x}_i$ a vector of length q for observations $i = 1, \ldots, n$.
- For *each observation* $i = 1, \ldots, n$, the vector is partitioned as

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ip} \\ x_{i1} \\ \vdots \\ x_{iq} \end{pmatrix}.$$

Sample mean vector and covariance matrix for subset of variables

- For the sample of n observation vectors, the mean vector and covariance matrix have the form

$$\begin{pmatrix} \overline{\mathbf{y}} \\ \overline{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \overline{y}_1 \\ \vdots \\ \overline{y}_p \\ \overline{x}_1 \\ \vdots \\ \overline{x}_q \end{pmatrix} \qquad \underset{(p+q)\times(p+q)}{\mathbf{S}} = \begin{pmatrix} \underset{p\times p}{\mathbf{S}_{yy}} & \underset{p\times q}{\mathbf{S}_{yx}} \\ \underset{q\times p}{\mathbf{S}_{xy}} & \underset{q\times q}{\mathbf{S}_{xx}} \end{pmatrix}$$

- Notice:
  (1) $\mathbf{S}$ is a symmetric matrix.
  (2) $\mathbf{S}_{xy} = \mathbf{S}'_{yx}$.

Sample mean vector and covariance matrix for subset of variables: Example

- Example: In the textbook, taken from Reaven and Miller (1979).
- Five variables: three main variables, and two secondary variables

$$
\begin{aligned}
x_1 &= \text{glucose intolerant} \\
x_2 &= \text{insulin response to oral glucose} \\
x_3 &= \text{insulin resistance} \\
y_1 &= \text{relative weight} \\
y_2 &= \text{fasting plasma glucose}
\end{aligned}
$$

y1	y2	x1	x2	x3
0.810	80	356	124	55
0.950	97	289	117	76
0.940	105	319	143	105
1.040	90	356	199	108
⋮	⋮	⋮	⋮	⋮

Sample mean vector and covariance matrix for subset of variables: Example

$$\begin{pmatrix} \bar{\mathbf{y}} \\ \bar{\mathbf{x}} \end{pmatrix} = \left(\begin{array}{c} \bar{y}_1 \\ \bar{y}_2 \\ \hline \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{array}\right) = \left(\begin{array}{r} 0.92 \\ 90.41 \\ \hline 340.83 \\ 171.37 \\ 97.78 \end{array}\right)$$

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix} = \left(\begin{array}{rr|rrr} 0.02 & 0.22 & 0.79 & -0.21 & 2.19 \\ 0.22 & 70.56 & 26.23 & -23.96 & -20.84 \\ \hline 0.79 & 26.23 & 1106.41 & 396.73 & 108.38 \\ -0.21 & -23.96 & 396.73 & 2381.88 & 1142.64 \\ 2.19 & -20.84 & 108.38 & 1142.64 & 2136.40 \end{array}\right)$$

More partitions are possible.

- Let $a_1, a_2, \ldots, a_p$ be known constants with $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$.
- Let $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$ be a single observation of p variables.
- Consider the linear combination

$$z = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p = \mathbf{a}'\mathbf{y}.$$

- Imagine now that the <u>same</u> known constant $\mathbf{a}$ is applied to *each* observation $\mathbf{y}_i$, $i = 1, \ldots, n$, giving

$$z_i = a_1 y_{i1} + a_2 y_{i2} + \cdots + a_p y_{ip} = \mathbf{a}'\mathbf{y}_i.$$

- What is the mean and variance of $z_1, z_2, \ldots, z_n$?
- Suppose you know the mean vector $\overline{\mathbf{y}}$ and covariance matrix $\mathbf{S}$ of $\mathbf{Y}$.

Linear combination of variables

- The sample mean $\overline{z}$ can be calculated as

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \mathbf{a}'\overline{\mathbf{y}}.$$

- The sample variance s_z^2 can be calculated as

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2 = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

- $\mathbf{a}'\mathbf{S}\mathbf{a}$ is the multivariate version of $s_z^2 = a^2 s_y^2$.
- $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$ is non-negative, for every $\mathbf{a}$.
  (Recall that if $\mathbf{S}$ is full rank ($n > p$), then $\mathbf{S}$ is positive definite.)

- The sample mean $\overline{z}$ can be calculated as

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \mathbf{a}'\overline{\mathbf{y}}.$$

  - Proof:
  - Recall $z_i = \mathbf{a}'\mathbf{y}_i$.
  - Then $\overline{z}$ satisfies

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}'\mathbf{y}_i = \mathbf{a}'\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i\right) = \mathbf{a}'\overline{\mathbf{y}}.$$

Linear combination of variables

- The sample variance s_z^2 can be calculated as

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

  - Proof:
  - $z_i = \mathbf{a}'\mathbf{y}_i$ and $\bar{z} = \mathbf{a}'\bar{\mathbf{y}}$ so that $z_i - \bar{z} = \mathbf{a}'\mathbf{y}_i - \mathbf{a}'\bar{\mathbf{y}} = \mathbf{a}'(\mathbf{y}_i - \bar{\mathbf{y}})$.
  - Being scalars we have $\mathbf{a}'(\mathbf{y}_i - \bar{\mathbf{y}}) = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{a}$.
  - This gives:

$$\begin{aligned}
\sum_i (z_i - \bar{z})^2 &= \sum_i \left\{\mathbf{a}'(\mathbf{y}_i - \bar{\mathbf{y}})\right\}^2 \\
&= \sum_i \left\{\mathbf{a}'(\mathbf{y}_i - \bar{\mathbf{y}})\right\}\left\{(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{a}\right\} = \sum_i \mathbf{a}'(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{a} \\
&= \mathbf{a}'\left\{\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'\right\}\mathbf{a} = (n-1)\mathbf{a}'\mathbf{S}\mathbf{a}.
\end{aligned}$$

Linear combination of variables

- Consider now a different linear transformation with $\mathbf{b}' = (b_1, b_2, \ldots, b_p)$, so $\mathbf{a} \neq \mathbf{b}$.
- The <u>same</u> known constant $\mathbf{b}$ is applied to *each* $\mathbf{y}_i$, $i = 1, \ldots, n$, giving

$$w_i = b_1 y_{i1} + b_2 y_{i2} + \ldots + b_p y_{ip} = \mathbf{b}'\mathbf{y}_i.$$

- As before $z_i = a_1 y_{i1} + a_2 y_{i2} + \ldots + a_p y_{ip} = \mathbf{a}'\mathbf{y}_i$.
- The sample covariance between z_i's and w_i's is given by

$$s_{zw} = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})(w_i - \overline{w}) = \mathbf{a}'\mathbf{S}\mathbf{b}.$$

- The sample correlation between z_i's and w_i's is given by

$$r_{zw} = \frac{s_{zw}}{\sqrt{s_z^2 s_w^2}} = \frac{\mathbf{a}'\mathbf{S}\mathbf{b}}{\sqrt{(\mathbf{a}'\mathbf{S}\mathbf{a})(\mathbf{b}'\mathbf{S}\mathbf{b})}}.$$

Linear combination of variables

- The sample covariance between z_i's and w_i's is given by

$$s_{zw} = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})(w_i - \overline{w}) = \mathbf{a}'\mathbf{S}\mathbf{b}.$$

  - Proof:
  - $z_i - \overline{z} = \mathbf{a}'\mathbf{y}_i - \mathbf{a}'\overline{\mathbf{y}} = \mathbf{a}'(\mathbf{y}_i - \overline{\mathbf{y}})$ and $w_i - \overline{w} = \mathbf{b}'\mathbf{y}_i - \mathbf{b}'\overline{\mathbf{y}} = \mathbf{b}'(\mathbf{y}_i - \overline{\mathbf{y}})$
  - Being scalars we have $\mathbf{b}'(\mathbf{y}_i - \overline{\mathbf{y}}) = (\mathbf{y}_i - \overline{\mathbf{y}})'\mathbf{b}$.
  - This gives:

$$\begin{aligned}
\sum_i (z_i - \overline{z})(w_i - \overline{w}) &= \sum_i \left\{\mathbf{a}'(\mathbf{y}_i - \overline{\mathbf{y}})\right\}\left\{\mathbf{b}'(\mathbf{y}_i - \overline{\mathbf{y}})\right\} \\
&= \sum_i \left\{\mathbf{a}'(\mathbf{y}_i - \overline{\mathbf{y}})\right\}\left\{(\mathbf{y}_i - \overline{\mathbf{y}})'\mathbf{b}\right\} \\
&= \mathbf{a}'\left\{\sum_i (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'\right\}\mathbf{b} = (n-1)\mathbf{a}'\mathbf{S}\mathbf{b}.
\end{aligned}$$

Linear combination of variables

- Now denote the constants $\mathbf{a}$ and $\mathbf{b}$ as $\mathbf{a}_1$ and $\mathbf{a}_2$, and write

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix}.$$

- For a single observation $\mathbf{y}$ define

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1'\mathbf{y} \\ \mathbf{a}_2'\mathbf{y} \end{pmatrix} = \mathbf{A}\mathbf{y}.$$

- $\mathbf{z}$ is a bivariate random variable.
- Suppose now we have observations $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ (each a p vector).
- For each observation $\mathbf{y}_i$, we have the linear combination $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i$, $i = 1, \dots, n$.
- Each $\mathbf{z}_i$ is a vector of size 2, and there are n of them.
- Let $\bar{\mathbf{z}}$ be a mean vector (of size 2) across $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$. Then

$$\bar{\mathbf{z}} = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix}.$$

Linear combination of variables

- Suppose we have $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i$, $i = 1, 2, \ldots, n$. $\mathbf{A}$ is a $(2 \times p)$ matrix.
- Sample mean vector $\bar{\mathbf{z}}$ can be obtained via

$$\bar{\mathbf{z}} = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1'\bar{\mathbf{y}} \\ \mathbf{a}_2'\bar{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix} \bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{y}}.$$

- Can use a similar construction for the sample covariance matrix of $\mathbf{z}$:

$$\begin{aligned} \mathbf{S}_z &= \begin{pmatrix} s_{z_1}^2 & s_{z_1 z_2} \\ s_{z_2 z_1} & s_{z_2}^2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 & \mathbf{a}_1'\mathbf{S}\mathbf{a}_2 \\ \mathbf{a}_2'\mathbf{S}\mathbf{a}_1 & \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix} \mathbf{S}(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{A}\mathbf{S}\mathbf{A}'. \end{aligned}$$

- This can be extended from the bivariate case to having k-variates. Thus let
$$\mathbf{z}' = (z_1, z_2, \ldots, z_k)$$
where $z_r = \mathbf{a}_r'\mathbf{y}$, $r = 1, 2 \ldots, k$.
- This extension only affects the size of $\mathbf{A}$.
- Suppose we have $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i$, $i = 1, 2, \ldots, n$. $\mathbf{A}$ is a $(k \times p)$ matrix.
- The principle remains the same:
$$\bar{\mathbf{z}} = \mathbf{A}\bar{\mathbf{y}}$$
$$\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}'.$$
- Notice that $\mathrm{tr}(\mathbf{A}\mathbf{S}\mathbf{A}') = \sum_{r=1}^{k} \mathbf{a}_r'\mathbf{S}\mathbf{a}_r$.
(To see this, look at the $k = 2$ case previously.)

Linear combination of variables: Example

- Timm (1975) reported response time (in ms) to ‘probe words’ in five positions in a sentence from 11 individuals

y_1	y_2	y_3	y_4	y_5
51	36	50	35	42
27	20	26	17	27
37	22	41	37	30
42	36	32	34	27
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Consider the linear combination

$$z = 3y_1 - 2y_2 + 4y_3 - y_4 + y_5 = (3, -2, 4, -1, 1)\mathbf{y} = \mathbf{a}'\mathbf{y}.$$

Linear combination of variables: Example

- Here $z_1 = \mathbf{a}'\mathbf{y}_1 = 288$; multiplication of **a** and the first *row* of the data matrix gives

$$z_1 = (3, -2, 4, -1, 1)\begin{pmatrix} 51 \\ 36 \\ 50 \\ 35 \\ 42 \end{pmatrix} = 288.$$

- Using the same principle gives $z_1, \ldots, z_{11}$ as

$$(288, 155, 224, 175, 192, 242, 236, 192, 173, 144, 146).$$

  The vector length is the same as the number of individuals/rows.
- This gives $\bar{z} = 197$ and $s_z^2 = 2084.00$.

Linear combination of variables: Example

- Notice

$$\bar{\mathbf{y}} = \begin{pmatrix} 36.09 \\ 25.55 \\ 34.09 \\ 27.27 \\ 30.73 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 65.09 & 33.65 & 47.59 & 36.77 & 25.43 \\ 33.65 & 46.07 & 28.95 & 40.34 & 28.36 \\ 47.59 & 28.95 & 60.69 & 37.37 & 41.13 \\ 36.77 & 40.34 & 37.37 & 62.82 & 31.68 \\ 25.43 & 28.36 & 41.13 & 31.68 & 58.22 \end{pmatrix}.$$

- The sample mean $\bar{z}$ can also be obtained as

$$\bar{z} = \mathbf{a}'\bar{\mathbf{y}} = (3, -2, 4, -1, 1) \begin{pmatrix} 36.09 \\ 25.55 \\ 34.09 \\ 27.27 \\ 30.73 \end{pmatrix} = 197.0$$

and the sample variance as $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a} = 2084.00.$

Sample covariance matrix again

- We defined the sample covariance matrix $\mathbf{S}$ as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' = \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - n\overline{\mathbf{y}}\overline{\mathbf{y}}'\right).$$

- In matrix notation: $\mathbf{S} = \frac{1}{n-1}\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$.
- We can similarly define

$$\mathbf{V} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' = \frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - n\overline{\mathbf{y}}\overline{\mathbf{y}}'\right).$$

- In matrix notation: $\mathbf{V} = \frac{1}{n}\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$.
- Notice $\mathbf{V} = \frac{n-1}{n}\mathbf{S}$.

Distance between vectors

- In a univariate setting, a difference between two quantities ('distance') is made meaningful by dividing the difference by its standard deviation, thus
$$\frac{|y_1 - y_2|}{\sigma} \quad \text{or} \quad \frac{|\bar{y} - \mu|}{\sigma_{\bar{y}}}.$$
- In multivariate sense, this is equivalent to defining the **squared** distance and standardising using the inverse of the covariance matrix, thus
$$d^2 = (\mathbf{y}_1 - \mathbf{y}_2)'\mathbf{S}^{-1}(\mathbf{y}_1 - \mathbf{y}_2)$$
or
$$D^2 = (\bar{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}).$$
- This gives the (squared) Mahalanobis distance.
- Some textbooks use **V** rather than **S** here.