# Multivariate statistics

Data

NULL SPACE

## 0.1 Organizing data

In multivariate analysis, we tend to analyze measurements made on several variables or features. We typically first select some $p$ variables to record or measure and then the values of these variables are recorded for each distinct observation or item. We denote $x_{jk}$ as the value taken by the $k^{th}$ variable observed for the $j^{th}$ item. So, if we have $n$ observations or items and for each of these we have $p$ measurements, we can easily package this set of observations into a $n \times p$ matrix as follows. Traditionally this is called a **data matrix**.

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{1}$$

### 0.1.1 Descriptive statistics

In order to summarize information contained in vast amounts of data, we use certain measures known as **descriptive statistics**. For instance, **sample mean** gives us the central tendency in the data and **sample variance** gives us a measure of spread of the data.

- **Sample mean** for the $k^{th}$ variable is given by:

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \tag{2}$$

- **Sample variance** for the $k^{th}$ variable is given by:

$$s_k^2 = s_{kk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k) \tag{3}$$

- Suppose now that we have a pair of measurements concerning two variables - where vector indicates one measurement of the two variables:

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \cdots, \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix} \tag{4}$$

Generally we say that $x_{j1}$ and $x_{j2}$ are the variable $1$ and variable $2$ measurements for the $j^{th}$ observation. Then a measure of linear association between the two variables is captured by the **sample covariance**:

$$s_{12} = \frac{1}{n-1}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2) \tag{5}$$

- The **sample Correlation coefficient** between two general variables $x_i$ and $x_k$ can be given by:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)}} \tag{6}$$

- When we have lots of variables and we want a convenient method to package all the **pairwise covariances and correlations** between all the $p$ variables, then we use the **sample covariance and correlation matrices** as follows in the respective equations:

$$\boldsymbol{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{7}$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{8}$$

The above two matrix are **square ($p \times p$) and symmetric matrices**.

## 0.2 Revising essential matrix algebra

- The **norm** or the length of a vector $\boldsymbol{x} = (x_1, x_2)$ is calulated by taking the square root of the dot product of the vector with itself ($\sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}$):

$$|\boldsymbol{x}| = \sqrt{x_1^2 + x_2^2} \tag{9}$$

We **normalize** a vector by dividing it by its length to obtain the corresponding **unit vector**. Usually when we compute **eigenvectors** - we usually normalize them and work with **unit eigenvectors**:

$$\boldsymbol{e} = \frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}} \tag{10}$$

- The **spectral** decomposition of a matrix $A$ is given below:

$$A = \lambda_1 \boldsymbol{e_1 e_1'} + \lambda_2 \boldsymbol{e_2 e_2'} + \cdots + \lambda_k \boldsymbol{e_k e_k'} \tag{11}$$

In general we say that a $p \times p$ matrix $A$ can be **diagonalized** or decomposed such that:

$$\boldsymbol{A = P \Lambda P'} \tag{12}$$

Where $\boldsymbol{P}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. Now for a $p \times p$ matrix there will be $p$ eigenvalues and eigenvectors. But the spectral theorem states that we can take the **most significant eigen values and vectors and recompute the compressed form of the original matrix without losing much information**. This basically means that we can choose $k < p$ such that we can recompute the original matrix as given in equation 11.

- A **quadratic form** is nothing but a quadratic equation written in matrix form as $\boldsymbol{x'Ax}$. The quadratic form is said to be **positive definite** if the following holds:

$$0 \leq \boldsymbol{x'Ax} \tag{13}$$

- An **illustrative example for quadratic forms**. Consider the following quadratic equation:

$$3x_1^2 + 2x_2^2 - 2\sqrt{2}x_1x_2 \tag{14}$$

This can be packaged into a matrix form as follows:

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 3 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{x'Ax} \tag{15}$$

Where $A$ is nothing but the **coefficient matrix** of the quadratic equation. We know that the eigenalues are solutions or roots to the characteristic equation given by:

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = (3 - \lambda)(2 - \lambda) - 2 = 0 \tag{16}$$

From the above example we find the eigenvalues to be $\lambda_1 = 4$ and $\lambda_2 = 1$. Using the **spectral decomposition** we can write:

$$\boldsymbol{A} = \lambda_1 \boldsymbol{e_1 e_1'} + \lambda_2 \boldsymbol{e_2 e_2'} = 4\boldsymbol{e_1 e_1'} + 1\boldsymbol{e_2 e_2'} \tag{17}$$

- A general notation of the spectral decomposition is given as:

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{e_i e_i'} \tag{18}$$

Where the $\boldsymbol{e_i}$ vectors are **normalized eigenvectors**. Now if we put all the normalized eigenvectors into a matrix as column vectors of the matrix called $\boldsymbol{P}$ such that:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{e_i} & \cdots & \boldsymbol{e_k} \end{bmatrix} \tag{19}$$

Then we can write the spectral decomposition in matrix form as:

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{e_i} \boldsymbol{e_i'} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P'} \tag{20}$$

Where the matrix of eigenvectors $\boldsymbol{P}$ is an **orthonormal matrix** that has the property that $\boldsymbol{P}\boldsymbol{P'} = \boldsymbol{P'}\boldsymbol{P} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \tag{21}$$

- We can denote $\boldsymbol{\Lambda}^{1/2}$ as the diagonal matrix with $\sqrt{\lambda_i}$ as its diagonal entries. Hence we can then write the **square root matrix of $\boldsymbol{A}$** as follows:

$$\boldsymbol{A}^{1/2} = \sum_{i=1}^{k} \sqrt{\lambda_i} \boldsymbol{e_i} \boldsymbol{e_i'} = \boldsymbol{P} \boldsymbol{\Lambda}^{1/2} \boldsymbol{P'} \tag{22}$$

## 0.3 The geometry of Data

First of all we note that since multivariate data is presented in the form of matrices and vector and these are nothing but geometric objects plotted in some coordinate space, we can get important insights from the data just by looking at the geometric properties of these data vectors and matrices. Let us have a good look at our data matrix that we presented in the start:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{23}$$

Each row of $\boldsymbol{X}$ represents a multivariate observation - we say that the entire dataset is a sample of $n$ size from a $p$-variate population - which in turn means that our data has $n$ measurements, each of which has $p$ components. If we think of all **observations as p dimensional vectors** then we can represent the data matrix as composed of $n$, $p$-dimensional row vectors:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1'} \\ \boldsymbol{x_2'} \\ \vdots \\ \boldsymbol{x_n'} \end{bmatrix} \tag{24}$$

Alternatively we can think of another geometrical interpretation of this data matrix by considering each column as a **feature vector** with $n$ components. Here each

column vector corresponds to **all the measurements for each variable**. So we will have $p$ column vector each of $n$ components. The data matrix then becomes:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{y_1} & \boldsymbol{y_2} & \cdots & \boldsymbol{y_p} \end{bmatrix} \tag{25}$$

Just to clarify, in the above matrix, $\boldsymbol{y_1} = [x_{11}, x_{21}, \cdots, x_{n1}]$ are the $n$ measurements on the first variable. Now we will look at a general geometric method of finding the sample mean.

- Let a vector of all 1s be represented as $\mathbf{1}$, where $\mathbf{1}_n$ is the $n$ component vector that contains only 1s - $[1, 1, \cdots, 1]$.
- Now to get the sample mean of the $i^{th}$ feature or variable we carry out the following computation:

$$\bar{x}_i = \frac{1}{n}\boldsymbol{y_i} \cdot \mathbf{1} = \frac{(x_{1i} + x_{2i} + \cdots + x_{ni})}{n} \tag{26}$$

  If we further multiply this vector with the $\mathbf{1}$ vector again, we would get a vector with all the values as $\bar{x}_i$:

$$\bar{x}_i\mathbf{1} = \begin{bmatrix} \bar{x}_i \\ \vdots \\ \bar{x}_i \end{bmatrix} \tag{27}$$

- To find the **deviation vector** we carry out the following manipulations:

$$\boldsymbol{d_i} = \boldsymbol{y_i} - \bar{x}_i\mathbf{1} = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix} \tag{28}$$

- Some general geometric matrix transformations are given below:

$$\textit{mean vector:} \quad \bar{\boldsymbol{x}} = \frac{1}{n}\boldsymbol{X'}\mathbf{1} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \tag{29}$$

$$\textit{mean matrix:} \quad \mathbf{1}\bar{\boldsymbol{x}}\boldsymbol{'} = \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X} = \begin{bmatrix} \bar{x}_1 & \cdots & \bar{x}_p \\ \vdots & \cdots & \vdots \\ \bar{x}_1 & \cdots & \bar{x}_p \end{bmatrix} \tag{30}$$

$$\textit{deviation matrix:} \quad \boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & \cdots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{bmatrix} \tag{31}$$

$$\textit{deviation covariance matrix:} \quad \boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X'}\left(\boldsymbol{I_n} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\boldsymbol{X} \tag{32}$$

# References

[1] Richard Arnold Johnson - Applied multivariate Statistical analysis