Mathematics and Statistics Brush-up: Linear Algebra and Descriptive Statistics

Albert Rodriguez Sala

UAB and Barcelona GSE

Linear Algebra

Scalars, vectors, and matrices

- A **scalar** a is a single number.
- A **vector** $\mathbf{a}$ is an element of Euclidean k space, written as $\mathbf{a} \in \mathbb{R}^k$.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

Scalars, vectors, and matrices

- A **matrix** is a $k \times r$ rectangular array of numbers, written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \vdots & & \ddots & \\ ak1 & a_{k2} & \dots & a_{kr} \end{bmatrix}$$

The transpose

The **transpose** of a matrix $A(k \times r)$ denoted A', A^r is obtained by flipping the matrix on its diagonal. If a matrix A is $k \times r$ then its transpose A' is $r \times k$. Properties of the transpose operator:

$$(A + B)' = A' + B'$$
$$(cA)' = cA'$$
$$(AB)' = B'A'$$

Squared, Symmetric, Diagonal and Identity matrices

- A matrix $A(k \times r)$ is **square** if $k = r$.
- A square matrix is **symmetric** if $A = A'$.
- A matrix is **diagonal** if the off-diagonal elements are all zero.
- An important diagonal matrix is the **identity matrix, or unit matrix**, $I_k(k \times k)$ which has ones on the diagonal. Let A be a $k \times r$ matrix, then:

$$AI_r = A$$
$$I_k A = A$$

Partitioned matrix

A **partitioned matrix** is a matrix such that its elements are matrices, vectors and/or scalars.

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1r} \\ A_{21} & A_{22} & \dots & A_{2r} \\ \vdots & & \ddots & \\ A_{k1} & A_{k2} & \dots & A_{kr} \end{bmatrix}$$

Matrix Addition

if matrices A, B are of the same order:

$$\underset{(k\times r)}{A} + \underset{(k\times r)}{B} = \begin{bmatrix} a_{11}+b_{11} & \dots & a_{1r}+b_{1r} \\ \vdots & \ddots & \\ ak1+b_{k1} & \dots & akr+b_{kr} \end{bmatrix}$$

Properties:

$$A+B=B+A(\text{Commutative law})$$
$$A+(B+C)=(A+B)+C(\text{associative law})$$

Matrix multiplication

Scalar multiplication: Let $c \in \mathbb{R}$ be a scalar and, A a matrix $(k \times r)$. The scalar product is defined st:

$$cA = Ac = (a_{ij}c) \qquad \forall ij \in k \times r.$$

Inner product (vectors): If **a** and **b** are both $k \times 1$, then their inner product is

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \ldots + a_k b_k = \sum_{j=1}^{k} a_j b_j$$

Note that $a'b = b'a$. We say that two vectors **a** and **b** are **orthogonal** if **a'b** $= 0$.

Matrix multiplication

Matrix product: If the number of columns in A is equal to the number of rows in B, then A and B are **conformable**. In this event the matrix product AB is defined as:

$$\underset{(k\times r)}{A}\underset{(r\times s)}{B} = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_k \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \dots & \mathbf{b}_k \end{bmatrix} = \begin{bmatrix} \mathbf{a_1}'\mathbf{b_1} & \dots & \mathbf{a_1}'\mathbf{b_s} \\ \vdots & \ddots & \\ \mathbf{a_k}'\mathbf{b_k} & \dots & \mathbf{a_k}'\mathbf{b_s} \end{bmatrix}$$

Matrix multiplication

Properties:

- Matrix multiplication is not commutative in general:

$$AB \neq BA$$

- Matrix multiplication is associative and distributive:

$$A(BC) = (AB)C$$
$$A(B + C) = AB + AC$$

the $(k \times r)$ Matrix A, $r \leq k$, is called orthonormal if $A'A = I_r$:

The trace and the rank

The **trace** of a square matrix A is the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{k} a_{ii}$$

The **rank** of the matrix $A(k \times r)$ with $r \leq k$ is the number of linearly independent columns $\mathbf{a}_j$ and is written as rank(A). We say A has full rank if rank(A)=r. A square matrix A is said to be **non-singular**, **invertible**, **non-degenerated** if it has full rank. This means that there is no vector $\mathbf{c}$ $(k \times 1)$, with $\mathbf{c} \neq 0$ such that $A\mathbf{c} = 0$.

Nonsingular matrices and the inverse

If a square matrix $(k \times k)$ A is nonsingular then there exists a unique matrix $(k \times k)$ A^{-1} called the **inverse** of A which satisfies:

$$AA^{-1} = A^{-1}A = I_k$$

Properties for non-singular matrices A and B include:

- $AA^{-1} = A^{-1}A = I_k$
- $(A^{-1})^{-1} = A$
- $(A^{-1})' = (A')^{-1}$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A+B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}$
- $det(A^{-1}) = det(A)^{-1}$

Also if A is an **orthonormal** matrix then $A^{-1} = A^{'}$.

The determinant

The **determinant** is a measure of the volume of a matrix $(k \times k)$. Let A be a (2×2) matrix, then its determinant is:

$$det(A) = |A| = a_{11}a_{22} - a_{12}a_{21}$$

Properties of the determinant:

- $det(A) = det(A')$
- $det(cA) = c^k det(A)$
- $det(AB) = det(A)det(B)$
- $det(A^{-1} = det(A)^{-1}$
- $det(A) \neq 0 \iff A$ is non-singular.

The characteristic equation and the eigenvalues of a matrix

The **characteristic equation** of a $(k \times k)$ matrix is:

$$det(A - \lambda I_k) = 0$$

It has k roots not necessarily distinct and real or complex. the **eigenvalues** or **latent roots** or **characteristic roots**; λ, of A are the roots of the characteristic function.

The eigenvector

If λ_i is an eigenvalue of A, then $A - \lambda_i I_k$ is singular so that there exists a non-zero vector $\mathbf{h_i}$ such that

$$(A - \lambda_i I_k)\mathbf{h_i} = 0$$

the vector h_i is called a **latent vector** or **characteristic vector** or **eigenvector** of A corresponding to λ_i. Let Λ be a diagonal matrix with the eigenvalues in the diagonal and let $H = [\mathbf{h_1} \dots \mathbf{h_k}]$. Some useful properties are:

Properties of eigenvalues and eigenvectors

- $det(A) = \prod_{i=1}^{k} \lambda_i$
- $tr(A) = \sum_{i=1}^{k} \lambda_i$
- A is non-singular $\iff \lambda_i \neq 0 \forall i = 1 \ldots k$
- if A has distinct eigenvalues, then there exists a nonsingular matrix P such that $A = P^{-1} \Lambda P$ and $PAP^{-1} = \Lambda$.
- If A is symmetric, then $A = H \Lambda H'$, which is called the **spectral decomposition** of A, and $H'AH = \Lambda$. In that case the eigenvalues are all real.
- the eigenvalues of A^{-1} are $\lambda_1^{-1}, \ldots, \lambda_k^{-1}$,
- The matrix H has the orthonormal properties: $H'H = I$, $HH' = I$, or alternatively, $H^{-1} = H'$ and $(H')^{-1} = H$

Note that if an eigenvalue has multiplicity 2, it can have 1 or 2 latent vectors linearly independent. If it has multiplicity 3, it can gave 1, 2, 3 latent vectors linearly independent.

Example

Exercise
Compute the eigenvalues and eigenvectors of the following matrices. Indicate clearly the multiplicity of each vector.

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Positive definiteness

A symmetric matrix is **positive (semi-)definite**, $A \underset{(\geq)}{>} 0$, if and only if $\forall \mathbf{c} \neq 0$:

$$\mathbf{c}' A \mathbf{c} \underset{(\geq)}{>} 0$$

A matrix A is larger than B if $\mathbf{c}'(A - B)\mathbf{c} > 0$.

Properties:

- $A > 0 \Longleftrightarrow \lambda_i \geq 0 \forall i$
- $A > 0 \longrightarrow$ A is non-singular and $A^{-1} \exists$ and $A^{-1} > 0$
- $A = G'G \longrightarrow A \geq 0$ and if G has full rank then, $A > 0$
- $A > 0 \longrightarrow \exists B \text{ st} A = BB'$ where B is **the matrix square root**.

Derivatives of matrices

Let $\mathbf{x} = [x_1 \dots x_k]$ be a $(k \times 1)$ vector and $g(\mathbf{x}) : \mathbf{R}^k \to \mathbf{R}$ a vector function. Then, the **vector derivative** is defined as:

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial g(\mathbf{x})}{\partial x_k} \end{bmatrix} \qquad \text{and} \qquad \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial g(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g(\mathbf{x})}{\partial x_k} \end{bmatrix}$$

Properties

Some properties:

$$\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}'\mathbf{a}}{\partial \mathbf{x}} = a$$

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = A$$

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (A + A')\mathbf{x}$$

$$\frac{\partial^2 \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}\partial \mathbf{x}'} = A + A'$$

Example: Deriving the OLS estimator in matrix notation

Exercise
Show that the OLS estimator in matrix notation is $\hat{\beta} = (X'X)^{-1}X'y$

Solution Take the linear model $y = X\beta + u$ where y is a $(n \times 1)$ vector of observations; X is a $(n \times k)$ matrix of regressors; β is a $(k \times 1)$ vector of coefficients; and u is a $(n \times 1)$ vector of errors. Then, the OLS estimator $\hat{\beta}$ is the vector of coefficients that minimize squared residuals:

$$\min_{\beta}\{u'u\} \qquad \{u'u\} = (y - X\beta)'(y - X\beta)$$

$$= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta$$

$$= y'y - 2\beta'X'y + \beta'X'X\beta$$

Example: Deriving the OLS estimator in matrix notation

note that the second and third term in the second line are scalars [1] and the fact that the transpose of a scalar is the scalar i.e. $y'X\beta = (y'X\beta)' = \beta'X'y$.

The FOC is:

$$\frac{\partial u'u}{\partial \beta} = \frac{\partial y'y - 2\beta'X'y + \beta'X'X\beta}{\partial \beta} = 0$$

[1] the size of the second term is $(1 \times n)(n \times k)(k \times 1) = (1 \times 1)$ while the size of the third is $(1 \times k)(k \times n)(n \times 1) = (1 \times 1)$

Example: Deriving the OLS estimator in matrix notation

Using matrix calculus properties (1) and (3):

$$\frac{\partial 2\beta' X' y}{\partial \beta} = 2X'y \qquad \frac{\partial \beta' X' X \beta}{\partial \beta} = 2X'X\beta$$

so that the FOC becomes:

$$\frac{\partial u'u}{\partial \beta} = -2X'y + 2X'X\beta = 0$$

Isolating β:

$$(X'X)\beta = X'y$$
$$(X'X)^{-1}(X'X)\beta = (X'X)^{-1}X'y$$

Hence, The OLS estimator is:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Descriptive Statistics

The statistical method

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. We typically work with a **statistical population** or **population** of interest, from which we associate a statistical model, a model under statistical assumptions concerning the generation of the data. Then, the **statistical method** is represented as follows

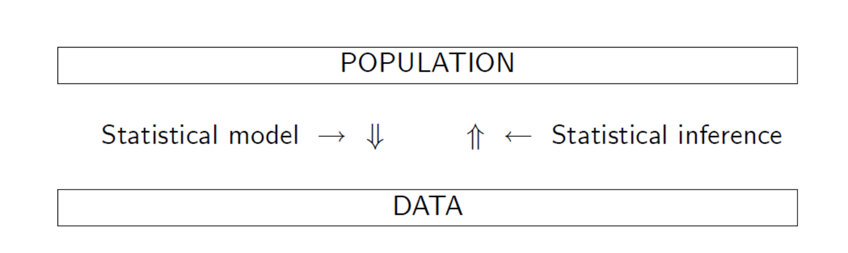


Figure 1: The statistical method (Figure from Jordi Caballé's lecture notes)

Statistical methods

Two main statistical methods are used in data analysis:

1. **descriptive statistics**: studies how to analyze and present the data.
2. **statistical inference**: by "inverting" the statistical model, allows us to say something (make inference) about the population from the data.

The statistical model is under the framework of **probability theory**, which deals with the analysis of random phenomena.

Descriptive statistics

A **population** is a set of people or objects of interests. the elements of a population are a called **individuals**. A **sample** is a subset of a population. Population and sample are relative concepts. A **variable** is a characteristic of a population which can take different values. Variables can be

1. **Qualitative (or categorical) variables**, which are the ones that cannot be measured numerically.
2. **Quantitative variables** which are the ones that can be measured numerically. Quantitative variables can be of 2 types:
   2.1 **Discrete** or **countable**, which are the ones that can take values from a countable set of numbers (like any list from the natural numbers $\mathbb{N}$). Discrete variables can be **finite** or **infinite**.
   2.2 **Continuous** which are the ones that can take values from an uncountable set of numbers (like the set of real numbers $\mathbb{R}$).

Relative frequencies

In descriptive statistics we apply indexes in one or more variable to describe the data. To start with, assume we care about one single variable X of the population. Then,

- The **absolute frequency** $n(x)$ of the value x is the number of times that the value appears in the data.
- The **relative frequency** $f(x)$ of the value x is the fraction or percentage of times that the value x appears in the data

$$f(x) = \frac{n(x)}{N}$$

Cumulative frequencies

- The **cumulative absolute frequency** $N(x)$ is the number of times that the variable X takes values smaller or equal than x

$$N(x) = \sum_{y \leq x} n(y) \tag{1}$$

- The **cumulative relative frequency** $F(x)$ is the fraction of times that the variable X takes values smaller or equal than x.

$$F(x) = \frac{N(x)}{N} = \frac{\sum_{y \leq x} n(y)}{N} = \sum_{y \leq x} \frac{n(y)}{N} = \sum_{y \leq x} f(y)$$

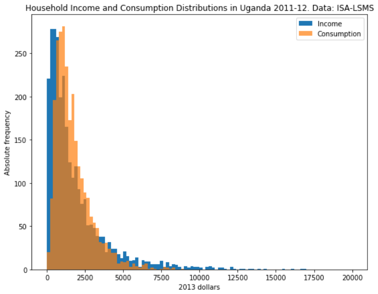
Histograms

Note that cumulative frequencies are only well-defined for quantitative (discrete or continuous) variables. In continuous variables and in discrete ones (with a "large" number of values) we partition the set of values into classes, intervals or bins. that is, we work with grouped data. When we work with grouped data we use **histograms** for the corresponding graphical representation.

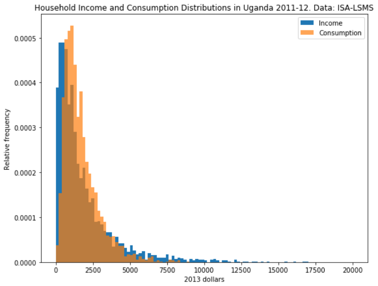
Histograms

Example

The histogram of income and consumption variables for households in Uganda in 2011/12. Data from the Uganda National Panel Survey (ISA-LSMS umbrella), a nationally representative sample.



(a) Absolute Frequency



(b) Relative Frequency

Measures of central tendency

the **Mean, average value or arithmetic mean** denoted $\overline{X}$ or $\overline{X}_n$, of a variable X is

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \sum_{x} x \cdot f(x)$$

where x_i is the value of the variable X for individual i. Properties of the mean, $\overline{X}$:

1. $\sum_{i=1}^{N}(x_i - \overline{X}) = 0$
2. $\overline{kX} = k\overline{X}$, where k is a constant.
3. Let $Z = \alpha X + \beta Y$, then:

$$\overline{Z} = \alpha\overline{X} + \beta\overline{Y}$$

Measures of central tendency

The **Median** is a value separating the higher half from the lower half of a data sample, a population or a probability distribution. To compute it, we order all values of the variable X taken by N individuals from the smallest to the largest, so that

$$x_1 \leq x_2 \leq \cdots \leq x_{N-1} \leq x_N$$

then, the median of X is

$$Median(X) = \begin{cases} x_{N/2+1/2} & \text{if } N \text{ is odd} \\ \frac{x_{N/2}+x_{N/2+1}}{2} & \text{if } N \text{ is even} \end{cases}$$

The **Mode** is the value that appears a larger number of times. The number with a larger absolute (and relative) frequency.

Measures of variability

The **variance** of a variable X, denoted by $Var(X)$ or S_X^2, measures the average of the square of the deviations from the mean

$$Var(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N}$$

Properties of the variance:

1. $Var(X) = \overline{X^2} - \overline{X}^2$
2. $Var(kX) = k^2 Var(X)$
3. Let $Z = \alpha X + \beta Y$, then:

$$Var(Z) = \alpha^2 Var(X) + \beta^2 Var(Y) + 2cov(X, Y)$$

Measures of variability

The **Standard deviation** denoted by $sd(X), S_X$ is $:= +\sqrt{Var(X)}$. Note that $sd(kX) = ksd(X)$ if $k \geq 0$.
The **coefficient of variation** is

$$CV(\overline{X}) = \frac{sd(X)}{|\overline{X}|} \qquad \text{when } \overline{X} \neq 0$$

Note that the coefficient of variation is *immune* to the units of measurement:

$$CV(kX) = \frac{sd(kX)}{|\overline{kX}|} = \frac{ksd(X)}{k|\overline{X}|} = CV(X) \quad \text{if } k > 0$$

Measures of variability

The **range** of the variable X is the diference between the largest and the smallest value taken by the variable. The **interquartile range** is the difference between the 75% value and the 25% one.

Other measures summarizing the shape of a distribution

The **Coefficient of skewness** or **Coefficient of asymmetry** is a measure of the asymmetry of the probability distribution of the variable X about its mean.

$$CA(X) = Skewness(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^3}{N \cdot sd(X)^3}$$

A negative value indicates that the tail is on the left side of the distribution, and positive a positive indicates that the tail is on the right. If the coefficient is zero, then the distribution of X is *symmetric*.

Other measures summarizing the shape of a distribution

The **coefficient of Kurtosis** is a measure of the thickness of the tails of the distribution and is given by

$$CK(X) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^4}{N \cdot sd(X)^4}$$

Multivariate frequency distributions

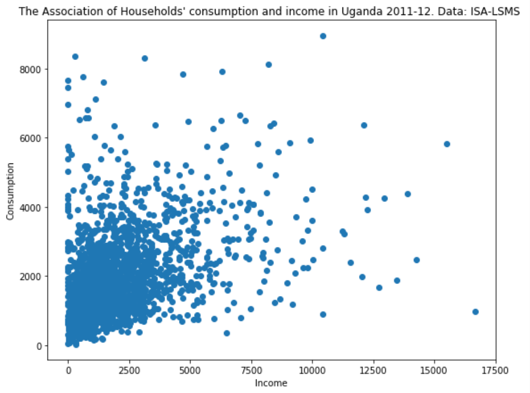
The multivariate frequency distribution or joint distribution gives us the distribution of several variables. For instance, the joint distribution of absolute frequencies of two variables X and Y gives us the number of times that each pair of values (x, y) corresponding to the pair of variables (X,Y) appears in the data.

Scatter plots

We can summarize these values in a **scatter plot**, which gives us an idea of the type of association or correlation between two quantitative variables.

Example

The scatter plot of consumption and income variables in Uganda 2011/12.



Multivariable frequencies

Similarly to the case of a single variable, we can define the relative frequency $f_{X,Y}(x,y)$ of the pair (x,y) as the fraction or percentage of times that this pair appears in the data:

$$f_{X,Y}(x,y) = \frac{n_{X,Y}(x,y)}{N}$$

Marginal frequency distributions

the **distribution of (absolute/relative) marginal frequencies of a variable** X is the frequency distribution of this variable with independency of the values taken by the other variables.

1. The **absolute marginal frequency** is

$$n_x(x) = \sum_y n_{X,Y}(x,y)$$

2. The **relative marginal frequency** $f_x(x)$ is

$$f_X(x) = \frac{n_X(x)}{N} = \sum_y f_{X,Y}(x,y)$$

Conditional frequencies distributions

The **distribution of conditional frequencies of the variable** X **given** $Y = y$ is the *relative* frequency distribution of the variable X for all the observations where $Y = y$.

- the **conditional frequency** $fX|Y(x, y)$ of the value x taken by the variable X given $Y = y$ is
$$fX|Y(x, y) = \frac{n_{X,Y}(x, y)}{n_Y(y)} = \frac{fX, Y(x, y)}{f_Y(y)}$$

The covariance

The **covariance** between X and Y, denoted $Cov(x, y)$ or $S_{X,Y}$, measures the strength of the association between these two variables and is given by

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{N}$$

Properties of the covariance:

1. $S_{X,X} = S_X^2$
2. $S_{X,Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$
3. $S_{\alpha X,\beta Y} = \alpha \cdot \beta \cdot S_{X,Y}$ where α and β are scalars.
4. $S_{X,Y} = S_{Y,X}$

The coefficient of correlation

The **coefficient of correlation**, denoted by $corr(X, Y)$ or ρ_X, Y, is given by

$$Corr(X, Y) = \frac{S_{X,Y}}{S_X \cdot S_Y}$$

Note that $Corr(\alpha X, \beta Y) = Corr(X, Y)$, so that the coefficient of correlation is *immune* to the units of measurement. Also note that $|Corr(X, Y)| \leq 1$.

Variables in matrix notation

When we study several variables it is useful to use the mean vector and the variance covariance matrix. Consider the following vector of K variables:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix}$$

Then, the **mean vector** or **vector of means** of variables $\mathbf{X}$ is

$$\overline{\mathbf{X}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix}$$

The **variance covariance matrix** or **covariance matrix** of the vector **X** of variables is the following $K \times K$ matrix

$$\mathbf{S} = \begin{bmatrix} S_1^2 & S_{12} & \dots & S_{1K} \\ S_{21} & S_2^2 & \dots & S_{2K} \\ \vdots & & \ddots & \\ S_{K1} & S_{K2}^2 & \dots & S_K^2 \end{bmatrix}$$

Note that since $\mathbf{S}_{ij} = \mathbf{S}_{ji}$ for all (i, j), the covariance matrix is symmetric.

Variables in matrix notation

Note that since $\mathbf{S}_{ij} = \mathbf{S}_{ji}$ for all (i, j), the covariance matrix is symmetric.
Now, let us define the vector of scalars $\alpha = [\alpha_1, \ldots, \alpha_K]'$ and define the variable $\mathbf{Z}$ as a linear combination of the variables appearing in $\mathbf{X}$: $\mathbf{Z} = \sum_j \alpha_j X_j = \alpha'\mathbf{X}$. Then,

1. $\overline{\mathbf{Z}} = \sum_j \alpha_j \overline{\mathbf{X}} = \alpha'\overline{\mathbf{X}}$
2. $\mathbf{S}_Z^2 = \sum_j \alpha_j^2 S_j^2 + 2\sum_i \sum_j \alpha_j \alpha_i S_i j = \alpha'\mathbf{S}\alpha$

Also note that $\mathbf{S^2_Z} = \alpha'\mathbf{S}\alpha \geq 0$ for all vector of scalars α. Therefore, the covariance matrix is positive semi-definite.