Raccogliamo su una popolazione di n individui i dati relativi a m caratteri (variabili) e riportiamoli in una matrice, dove le righe (n) sono relative ad individui diversi e le colonne (m) sono relative a caratteri diversi.

$$\mathbb{X} := \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \tag{3.11}$$

La costruzione della matrice ha senso, ovviamente, quando prendiamo in considerazione solo variabili numeriche.

Esempi. Peso, altezza, età di un gruppo di persone; peso, dimensioni, porosità di un campione di laterizi, etc...

È evidente che per ciascuna variabile (ovvero, per ogni colonna j) ha senso calcolare la media, la varianza e la deviazione standard:

$$\mu_j := \bar{x}_j = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} x_{ij} \,;$$

$$\sigma_j^2 := Var[x_j] = \tfrac{1}{n-1} \textstyle\sum_{i=1}^{n} (x_{ij} - \mu_j)^2 \,;$$

$$\sigma_j := Std[x_j] = \sqrt{Var[x_j]} \,,$$

Se indichiamo con X_h la variabile corrispondente all'$h-$esima colonna della matrice (3.11), possiamo definire la covarianza fra due caratteri qualunque (diciamo il $j-$esimo e l'$\ell-$esimo):

$$Cov[X_j, X_\ell] := \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \mu_j)(x_{i\ell} - \mu_\ell) \,.$$

La definizione implica ovviamente che

$$Cov[X_j, X_j] = Var[X_j] \,.$$

Qundi, ad esempio

$$c_{11} := Cov[X_1, X_1] = Var[X_1] = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \mu_1)^2$$

e per la simmetria della covarianza, i.e.

$$Cov[X_j, X_\ell] = Cov[X_\ell, X_j]$$

avremo, ad esempio

$$c_{12} := Cov[X_1, X_2] = c_{21} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \mu_1)(x_{i2} - \mu_2)\,.$$

La matrice costruita con le covarianze c_{ij} è dunque una matrice $(k \times k)$, simmetrica, dove sulla diagonale compaiono le varianze dei caratteri in esame:

$$c_{jj} = Cov[X_j, X_j] = Var[X_j, X_j]\,.$$

$$\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{12} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ c_{k1} & c_{k2} & \cdots & x_{kk} \end{pmatrix} \tag{3.12}$$

La matrice dei dati X può essere riportata ad una versione **standardizzata**, in modo da rendere le variabili adimensionali e a media nulla:

$$x_{ij} \longrightarrow y_{ij} := \frac{x_{ij} - \bar{x}_j}{\sigma_j}\,,$$

dove σ_j è la deviazione standard della variabile $j-$esima. La matrice Y sostituisce la X e i dati standardizzati Y_J sostituiscono i dati reali X_j, $j = 1, 2, \ldots, k$. Verifichiamo che gli Y_J abbiano media nulla e varianza unitaria:

$$\bar{Y_j} := \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij} - \mu_j}{\sigma_j} = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{x_{ij}}{\sigma_j} - n\frac{\mu_j}{\sigma_j}\right) = 0\,.$$

$$\sigma_j^2 = Var[Y_j] := \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \bar{Y_j})^2 = \frac{1}{n-1}\sum_{i=1}^{n} y_{ij}^2$$

poiché $\bar{Y_j} = 0$. Avremo quindi

$$Var[Y_j] = \frac{1}{n-1}\sum_{i=1}^{n} y_{ij}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_{ij} - \mu_j)^2}{\sigma_j^2} = \frac{1}{(n-1)\sigma_j^2}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2 = \frac{(n-1)\sigma_j^2}{(n-1)\sigma_j^2} = 1\,.$$

Dalla matrice dei dati standardizzati Y, possiamo costruire la matrice di covarianza $Cov[Y_h, Y_\ell]$, che risulterà essere una matrice quadrata e simmetrica di dimensioni $k \times k$ (k è il numero di variabili misurate sugli n individui).
Esempio

Consideriamo una matrice di covarianza per le variabili standardizzate Y_h (h è l'$h-$simo carattere, h=1,2,...,k.), vettori di dimensione n, pari al numero di individui del campione:

$$C_Y := Cov[Y_h, Y_m]\,, h, m = 1, 2, \ldots, k.$$

Se indichiamo con Y (e con Y^t la sua trasposta) la matrice $(n \times k)$ dei dati standardizzati, e quindi a media nulla, è facile verificare che

$$C_Y = \frac{1}{n-1} Y^t Y\,.$$

Costruiamo, a titolo di esempio, una matrice C_Y ad hoc, una matrice di piccole dimensioni (3×3), che abbia un paio di autovalori grandi:

$$C_Y = \begin{pmatrix} 6.3750 & -1.5155 & -2.2500 \\ -1.5155 & 8.1250 & -1.2990 \\ -2.2500 & -1.2990 & 2.500 \end{pmatrix} \tag{3.13}$$

Utilizzando Matlab, con l'istruzione $eig(C_Y)$, otteniamo gli autovalori di C_Y :

$$eig(C_Y) = (9, 7, 1)\,,$$

con due autovalori molto maggiori del terzo. Cerchiamo adesso una matrice U, ortogonale, che faccia passare dalle variabili standardizzate, ma fra loro correlate, Y (matrice $n \times k$), alla matrice di variabili Z, delle stesse dimensioni, ma di variabili standardizzate e incorrelate. Cerchiamo quindi di trovare delle combinazioni lineari delle Y che producano delle nuove variabili Z incorrelate:

$$Z = YU \quad \Longleftrightarrow \quad Z^t = U^t Y^t\,. \tag{3.14}$$

Le nuove variabili Z, in quanto combinazione lineare delle variabili standardizzate Y hanno ancora media nulla. La matrice di covarianza per i vettori $Z_\ell\,, \ell = 1, 2, \ldots, k\,,$ è ancora esprimibile come

$$C_Z := Cov[Z_i, Z_j] = \frac{1}{n-1} Z^t Z\,.$$

Abbiamo quindi, usando la (3.14):

$$C_Z = \frac{1}{n-1} Z^t Z = \frac{1}{n-1} Z^t Y U = \frac{1}{n-1} U^t Y^t Y U = U^t C_Y U\,. \tag{3.15}$$

Consideriamo la seguente matrice ortogonale:

$$U = \begin{pmatrix} 0.5000 & 0.7500 & 0.4330 \\ -0.8660 & 0.4330 & 0.2500 \\ 0 & -0.5000 & 0.8660 \end{pmatrix} \tag{3.16}$$

U è stata ottenuta come prodotto di due matrici di rotazione attorno a due assi non ortogonali (vedere gli angoli di Eulero).

$$U = U_1 * U_2$$

dove

$$U_1 = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.500 & 0.866 & 0 \\ -0.866 & 0.500 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

e

$$U_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{\sqrt{3}}{2} & \frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.866 & 0.500 \\ 0 & -0.500 & 0.866 \end{pmatrix}$$

Osserviamo che, sempre utilizzando Matlab,

$$\det U = 1\,.$$

Se diagonalizziamo C_Y, otteniamo:

$$C_Z = U^t * C_Y * U = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.17}$$

Possiamo adesso vedere come risulta la composizione di$Z$ in termini delle Y.

La matrice dei dati standardizzati e incorrelati Z si ottiene dalla matrice dei dati standardizzati Y per mezzo del prodotto con la matrice ortogonale U:

$$Z = Y * U$$

Vediamo che tipo di matrice di dati può generare una matrice di covarianza 3×3 e consideriamo una matrice di dati 2×3, ovvero una matrice dove sono rilevati tre caratteri diversi su due individui:

$$Z = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ & & \\ z_{21} & z_{22} & z_{23} \end{pmatrix}$$

Analogamente per la matrice Y avremo

$$Y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ \\ y_{21} & y_{22} & y_{23} \end{pmatrix}$$

Per effettuare il calcolo $Y * U$ con più semplicità possiamo scrivere U, definita in (3.16), come

$$U = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ \\ u_{21} & u_{22} & u_{23} \\ \\ u_{31} & u_{32} & u_{33} \end{pmatrix}$$

La variabile Z_1 è quella che ha varianza più grande (9) e corrisponde alla prima colonna della matrice Z:

$$Z_1 = \begin{pmatrix} z_{11} \\ \\ z_{21} \end{pmatrix}$$

Calcolando esplicitamente

$$\begin{cases} z_{11} = u_{11}y_{11} + u_{21}y_{12} + u_{31}y_{13} \\ \\ z_{21} = u_{11}y_{21} + u_{21}y_{22} + u_{31}y_{23} \end{cases}$$

che, riassunta in termini vettoriali dà:

$$Z_1 = u_{11}Y_1 + u_{21}Y_2 + u_{31}Y_3 = 0.5Y_1 - 0.866Y_2\,.$$

Procedendo in modo del tutto analogo per

$$Z_2 = \begin{pmatrix} z_{12} \\ \\ z_{22} \end{pmatrix}$$

abbiamo

$$Z_2 = u_{12}Y_1 + u_{22}Y_2 + u_{32}Y_3 = 0.75Y_1 + 0.433Y_2 - 0.5Y_3\,.$$