CdL in Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Lezione : Varianza totale e generalizzata

Docente: Aldo Solari

Nel caso $p = 1$, la variabilità (o dispersione) presente nelle misurazioni della variabile considerata è descritta da un singolo numero, la varianza $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Nel caso $p > 1$, la variabilità presente nelle misurazioni delle p variabili considerate è descritta da p varianze $s_{jj}, j = 1, \ldots, p$ e $\frac{1}{2}p(p-1)$ covarianze $s_{jk}, j \neq k = 1, \ldots, p$, ovvero

$$p + \frac{1}{2}p(p-1)$$

numeri, contenuti nella matrice di varianze/covarianze $\underset{p\times p}{S}$ Possiamo riassumere la variabilità descritta da $\underset{p\times p}{S}$ in un singolo numero (senza perdere troppa informazione)?

1 Varianza totale

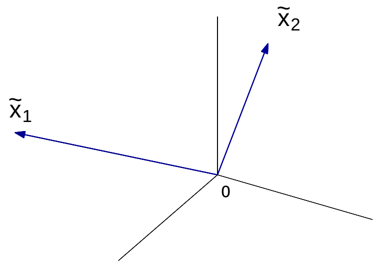
$$\boxed{\text{Varianza totale} = \mathrm{tr}(\underset{p\times p}{S}) = \sum_{j=1}^{p} s_{jj}}$$

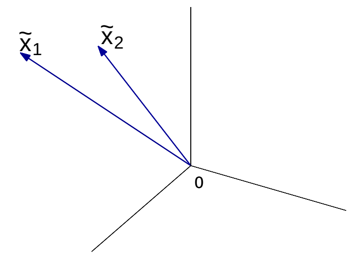
1.1 $\mathrm{tr}(S)$ nello spazio delle osservazioni

Nello spazio delle osservazioni, la varianza totale può essere interpretata come la somma delle lunghezze al quadrato dei p vettori scarto dalla media $\underset{p\times 1}{\tilde{x}_j}$, $j = 1, \ldots, p$, divisa per n.

$$\mathrm{tr}(\underset{p\times p}{S}) = \frac{1}{n}\sum_{j=1}^{p} n s_{jj} = \frac{1}{n}\sum_{j=1}^{p} \| \underset{p\times 1}{\tilde{x}_j} \|^2$$

Sintetizzando la matrice di varianze/covarianze con un singolo numero dato dalla varianza totale, perdiamo tutta l'informazione sulla struttura di correlazione (di covarianza) tra le p variabili.





$$\underset{3\times 2}{\tilde{X}} = \begin{bmatrix} 2 & -2 \\ -3 & 0 \\ 1 & 2 \end{bmatrix} \quad \mathrm{tr}(\underset{2\times 2}{S}) = \frac{14}{3} + \frac{8}{3} \quad r_{12} = -0.19$$

$$\underset{3\times 2}{\tilde{X}} = \begin{bmatrix} 1 & 0 \\ -3 & -2 \\ 2 & 2 \end{bmatrix} \quad \mathrm{tr}(\underset{2\times 2}{S}) = \frac{14}{3} + \frac{8}{3} \quad r_{12} = 0.95$$

1.2 $\mathrm{tr}(S)$ nello spazio delle variabili

La distanza Euclidea

- tra due unità statistiche $\underset{1\times p}{u'_i}$ e $\underset{1\times p}{u'_l}$:

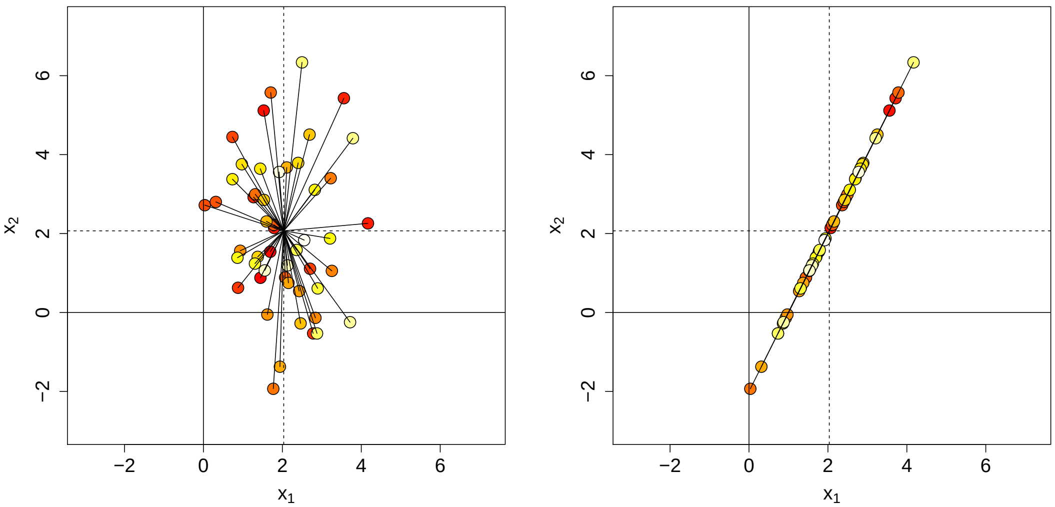
$$d(u_i, u_l) = \sqrt{(\underset{1\times p}{u_i - u_l})'(\underset{p\times 1}{u_i - u_l})} = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{lj})^2}$$

- tra l'i-sima unità statistica $\underset{1\times p}{u'_i}$ e il baricentro $\underset{1\times p}{\bar{x}'}$:

$$d(u_i, \bar{x}) = \sqrt{(\underset{1\times p}{u_i - \bar{x}})'(\underset{p\times 1}{u_i - \bar{x}})} = \sqrt{\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2}$$

Nello spazio delle variabili, la varianza totale può essere interpretata come la media aritmetica delle distanze Euclidee al quadrato delle n unità statistiche $\underset{1\times p}{u_i'}$, $i = 1, \ldots, n$, dal baricentro $\underset{1\times p}{\bar{x}'}$

$$\mathrm{tr}(\underset{p\times p}{S}) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2 = \frac{1}{n} \sum_{i=1}^{n} d^2(u_i, \bar{x})$$

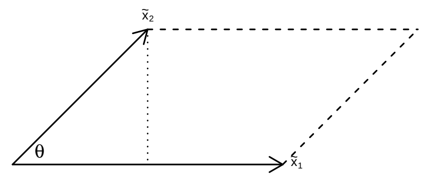


2 Varianza generalizzata

$$\boxed{\text{Varianza generalizzata} = \det(\underset{p\times p}{S})}$$

2.1 $\det(S)$ nello spazio delle osservazioni

Consideriamo geometricamente l'area generata da $p = 2$ vettori scarto dalla media $\underset{n\times 1}{\tilde{x}_1}$ e $\underset{n\times 1}{\tilde{x}_2}$ nello



spazio n-dimensionale

$$\begin{aligned}
\text{Area parallelogramma} &= \text{base paral.} \cdot \text{altezza paral.} \\
&= \parallel \underset{n\times 1}{\tilde{x}_1} \parallel \cdot \parallel \underset{n\times 1}{\tilde{x}_2} \parallel \sqrt{1 - \cos^2(\theta)} \\
&= n\sqrt{s_{11}s_{22}(1 - r_{12}^2)}
\end{aligned}$$

$$\begin{aligned}
\det(\underset{2\times 2}{S}) &= \det\left(\begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}r_{12} & s_{22} \end{bmatrix}\right) \\
&= s_{11}s_{22} - s_{11}s_{22}r_{12}^2 \\
&= s_{11}s_{22}(1 - r_{12}^2)
\end{aligned}$$

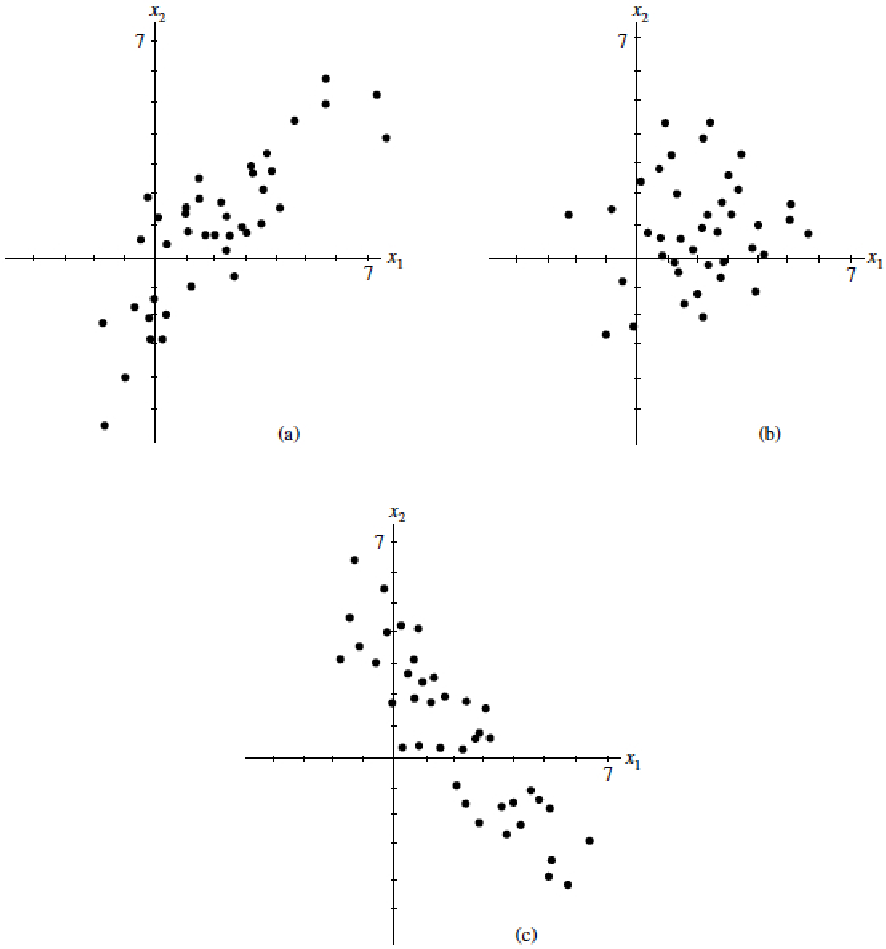
Quindi

$$\det(\underset{2\times 2}{S}) = \frac{(\text{Area parallelogramma})^2}{n^2}$$

In generale, per p vettori n-dimensionali $\underset{n\times 1}{\tilde{x}_j}$, $j = 1, \ldots, p$:

$$\det(\underset{p\times p}{S}) = \frac{(\text{Volume parallelepipedo } p\text{−dimensionale})^2}{n^p}$$

2.2 $\det(S)$ nello spazio delle variabili



Alla matrice di varianze/covarianze $\underset{p\times p}{S}$ possiamo associare p coppie di autovalori e autovettori

$$(\lambda_1, \underset{p\times 1}{v_1}), (\lambda_2, \underset{p\times 1}{v_2}), \ldots, (\lambda_p, \underset{p\times 1}{v_p})$$

dove gli autovalori (*eigenvalues*) sono ordinati in maniera decrescente, ovvero

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

e dove gli autovettori (*eigenvectors*) $\underset{p\times 1}{v_1}, \ldots, \underset{p\times 1}{v_p}$ sono tali che

- hanno lunghezza unitaria $\| \underset{1\times p}{v_1} \| = \ldots = \| \underset{1\times p}{v_p} \| = 1$
- sono mutualmente perpendicolari: $\langle v_k, v_j \rangle = \underset{1\times p}{v_j'} \underset{p\times 1}{v_k} = 0$ per $j \neq k$

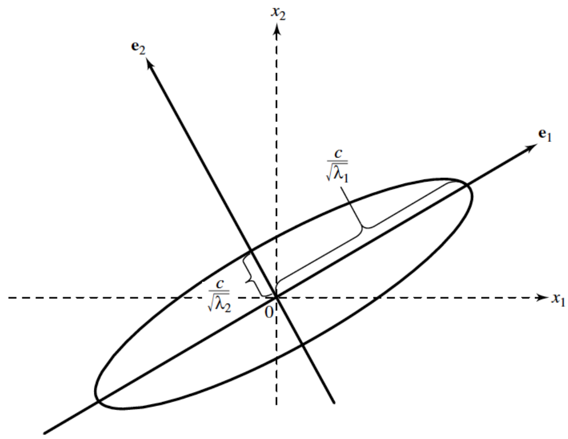


Figure 2.6 Points a constant distance c from the origin ($p = 2, 1 \leq \lambda_1 < \lambda_2$).

L'equazione

$$\underset{1\times p}{(x-\bar{x})'}\underset{p\times p}{S^{-1}}\underset{p\times 1}{(x-\bar{x})} = c^2$$

definisce l'iper-elissoide

- centrato sul baricentro $\underset{1\times p}{\bar{x}'}$
- con il j-simo asse orientato secondo il j-simo autovettore $\underset{p\times 1}{v_j}$ di $\underset{p\times p}{S}$
- di lunghezza $\dfrac{c}{\sqrt{\lambda_j}}$, proporzionale al j-simo autovalore λ_j di $\underset{p\times p}{S}$,

dove stiamo assumendo che $\underset{p\times p}{S}$ è una matrice definita positiva in modo da garantire l'esistenza di $\underset{p\times p}{S^{-1}}$ Il volume dell'iper-ellissoide è funzione della varianza generalizzata:

$$\text{Volume di } \left\{ \underset{1\times p}{x'} : \underset{1\times p}{(x-\bar{x})'}\underset{p\times p}{S^{-1}}\underset{p\times 1}{(x-\bar{x})} \leq c^2 \right\} = k_p c^p \sqrt{\det(S)}$$

dove $k_p = \dfrac{2\pi^{p/2}}{p\Gamma(p/2)}$ e $\Gamma(\cdot)$ è la funzione Gamma. Quindi

$$(\text{Volume iperellissoide})^2 = (\text{costante})(\text{varianza generalizzata})$$

Varianza generalizzata: cosa perdiamo: Sintetizzando la matrice di varianze/covarianze con un singolo numero dato dalla varianza generalizzata, perdiamo l'informazione riguardante l'orientamento della nuvola di punti p-dimensionale formata dalle n unità statistiche

2.3 Quando la varianza generalizzata è zero?

Proposition 2.1. *La varianza generalizzata è 0 se e solo se le colonne di $\underset{n\times p}{\tilde{X}}$ sono linearmente dipendenti.*

Dimostrazione. Si ricordi che le colonne di $\underset{n\times p}{\tilde{X}}$, ovvero i vettori $\underset{n\times 1}{\tilde{x}_j}$, $j = 1, \ldots, p$, sono linearmente dipendenti se esiste un vettore non nullo $\underset{p\times 1}{c} \neq \underset{p\times 1}{0}$ tale che

$$\underset{n\times p}{\tilde{X}}\underset{p\times 1}{c} = c_1 \underset{n\times 1}{\tilde{x}_1} + \ldots + c_p \underset{n\times 1}{\tilde{x}_p} = \underset{n\times 1}{0}$$

$\Leftarrow$

Se le colonne di $\underset{n\times p}{\tilde{X}}$ sono linearmente dipendenti, esiste $\underset{p\times 1}{c} \neq \underset{p\times 1}{0}$ tale che

$$\underset{n\times 1}{0} = \underset{n\times p}{\tilde{X}}\underset{p\times 1}{c}$$

quindi

$$n \underset{p\times p}{S}\underset{p\times 1}{c} = \underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}}\underset{p\times 1}{c} = \underset{p\times n}{\tilde{X}'}\underset{n\times 1}{0} = \underset{p\times 1}{0}.$$

Segue che esiste $\underset{p\times 1}{c} \neq \underset{p\times 1}{0}$ tale che $\underset{p\times p}{S}\underset{p\times 1}{c} = \underset{p\times 1}{0}$, ovvero che $\underset{p\times p}{S}$ è una matrice singolare, e quindi $\det(\underset{p\times p}{S}) = 0$

$\Rightarrow$

Se $\det(\underset{p\times p}{S}) = 0$, allora $\underset{p\times p}{S}$ è singolare ed esiste $\underset{p\times 1}{c} \neq \underset{p\times 1}{0}$ tale che $\underset{p\times p}{S}\underset{p\times 1}{c} = \underset{p\times 1}{0}$, ovvero

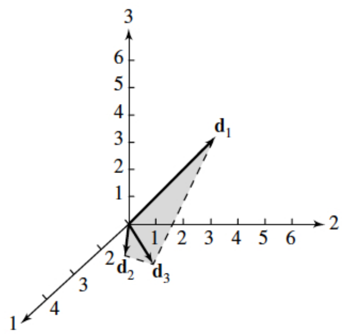
$$\begin{aligned}
\underset{p\times 1}{0} &= n \underset{p\times p}{S}\underset{p\times 1}{c} = \underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}}\underset{p\times 1}{c} \\
\underset{1\times p}{c'}\underset{p\times 1}{0} &= \underset{1\times p}{c'}\underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}}\underset{p\times 1}{c} \\
0 &= \| \underset{n\times p}{\tilde{X}}\underset{p\times 1}{c} \|^2
\end{aligned}$$

e quindi per avere lunghezza 0 dobbiamo avere $\underset{n\times p}{\tilde{X}}\underset{p\times 1}{c} = \underset{n\times 1}{0}$, ovvero le colonne di $\underset{n\times p}{\tilde{X}}$ sono linearmente dipendenti. $\square$

Example 2.2. $\underset{3\times 3}{X} = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{bmatrix}$, $\underset{3\times 3}{\tilde{X}} = \begin{bmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix}$*, quindi poichè*

$$\underset{3\times 1}{\tilde{x}_3} = \underset{3\times 1}{\tilde{x}_1} + 2 \underset{3\times 1}{\tilde{x}_2}$$

le colonne $\underset{3\times 3}{\tilde{X}}$ sono linearmente dipendenti, ovvero $\underset{3\times 3}{\tilde{X}}\underset{3\times 1}{c} = \underset{3\times 1}{0}$ con $\underset{3\times 1}{c} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} \neq \underset{3\times 1}{0}$. Geometricamente questo significa che uno dei vettori scarto dalla media, ad esempio $\underset{3\times 1}{\tilde{x}_3}$, giace *nel piano generato da $\underset{3\times 1}{\tilde{x}_1}$ e $\underset{3\times 1}{\tilde{x}_2}$. Di conseguenza, il volume del parallelepipedo tridimensionale è 0.*



Proposition 2.3. *Se* $n \leq p$, *allora* $\det(S) = 0$

Dimostrazione. Sia $\underset{1\times p}{\tilde{u}_i'} = \begin{bmatrix} \tilde{x}_{i1} & \cdots & \tilde{x}_{ip} \end{bmatrix}$ l'i-sima riga di $\underset{n\times p}{\tilde{X}}$. Abbiamo

$$\underset{p\times n}{\tilde{X}'}\underset{n\times 1}{1} = 1 \cdot \underset{p\times 1}{\tilde{u}_1} + \ldots + 1 \cdot \underset{p\times 1}{\tilde{u}_n} = \begin{bmatrix} \sum_{i=i}^n \tilde{x}_{i1} \\ \vdots \\ \sum_{i=i}^n \tilde{x}_{ip} \end{bmatrix} = \underset{p\times 1}{0}$$

quindi le righe di $\underset{n\times p}{\tilde{X}}$ sono linearmente dipendenti. Allora $\text{rango}(\underset{n\times p}{\tilde{X}}) < n \leq p$. Segue che

$$\text{rango}(\underset{n\times p}{\tilde{X}}) = \text{rango}(\underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}}) = \text{rango}(n\underset{p\times p}{S}) = \text{rango}(\underset{p\times p}{S}) < p$$

e quindi $\underset{p\times p}{S}$ è singolare, e risulta $\det(\underset{p\times p}{S}) = 0$ □

2.4 Varianza generalizzata per dati standardizzati

$$\boxed{\begin{array}{c}\text{Varianza generalizzata per} \\ \text{dati standardizzati } \underset{n\times p}{Z}\end{array} = \det(\underset{p\times p}{S^Z}) = \det(\underset{p\times p}{R})}$$

$$\begin{aligned} \det(\underset{p\times p}{S}) &= \det(\underset{p\times p}{D^{1/2}}\underset{p\times p}{R}\underset{p\times p}{D^{1/2}}) \\ &= \det(\underset{p\times p}{D^{1/2}})\det(\underset{p\times p}{R})\det(\underset{p\times p}{D^{1/2}}) \\ &= (s_{11}s_{22}\cdots s_{pp})\det(\underset{p\times p}{R}) \\ &= \left(\prod_{j=1}^p s_{jj}\right)\det(\underset{p\times p}{R}) \end{aligned}$$

dove $\underset{p\times p}{D^{1/2}} = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$

Example 2.4. *Se cambiamo l'unità di misura per la prima variabile* $\underset{n\times 1}{x_1}$*, ad esempio da Kg a gr, e quindi moltiplicando* $\underset{n\times 1}{x_1}$ *per 1000, abbiamo che la varianza* s_{11} *aumenta di un fattore moltiplicativo pari a* 1000^2*. Questo cambio di unità di misura da Kg a gr influenza la varianza generalizzata:*

$$\det(\underset{p\times p}{S^{gr}}) = \big((1000^2 s_{11})s_{22}\cdots s_{pp}\big)\det(\underset{p\times p}{R}) = 1000^2\det(\underset{p\times p}{S^{Kg}})$$

Per questo motivo, spesso è conveniente calcolare la varianza generalizzata considerando i dati standardizzati $\underset{n\times p}{Z}$

2.5 Indice relativo di variabilità

$$0 \leq \begin{array}{c}\text{Indice relativo}\\ \text{di variabilita}'\end{array} = \det(\underset{p\times p}{R}) = \frac{\det(\underset{p\times p}{S})}{\prod_{j=1}^{p} s_{jj}} \leq 1$$