4.1 Covarianze e coefficienti di correlazione

Definizione 4.1.1. La *varianza* di una variabile aleatoria reale X è

$$\mathrm{Var}(X) \stackrel{\mathrm{def}}{=} \mathbb{E}[(X-\mu_X)^2]$$

dove $\mu_X \stackrel{\mathrm{def}}{=} \mathbb{E}\, X$.

Definizione 4.1.2. La *covarianza* tra due variabili aleatorie reali X e Y è

$$\mathrm{Cov}(X,Y) \stackrel{\mathrm{def}}{=} \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

dove $\mu_X \stackrel{\mathrm{def}}{=} \mathbb{E}\, X$, $\mu_Y \stackrel{\mathrm{def}}{=} \mathbb{E}\, Y$ sono i valori attesi delle variabili.

La varianza di X può essere scritta come covarianza tra X e se stessa: $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.

Proposizione 4.1.1. *Siano X, Y, Z variabili aleatorie reali. La covarianza soddisfa le seguenti proprietà:*

1. $\mathrm{Cov}(X,X) \geq 0$;
2. $\mathrm{Cov}(X,X) = 0$ *se e solo se esiste* $\mu \in \mathbb{R}$ *tale che* $\Pr[X=\mu]=1$;
3. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$;
4. $\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y)$ *per ogni* $a \in \mathbb{R}$;
5. $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.

Si noti in particolare che la funzione $(X,Y) \mapsto \mathrm{Cov}(X,Y)$ soddisfa gli assiomi di un prodotto scalare (in uno spazio vettoriale in cui identifichiamo due variabili aleatorie X ed Y ogniqualvolta $\Pr[X-\mu_X = Y-\mu_Y] = 1$).

Definizione 4.1.3. Il *coefficiente di correlazione* tra due variabili aleatorie reali X e Y è definito come

$$\rho_{X,Y} \stackrel{\mathrm{def}}{=} \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Teorema 4.1.2. *Il coefficiente di correlazione tra X ed Y soddisfa $-1 \leq \rho_{X,Y} \leq 1$.*

Dimostrazione. Sia $\langle X,Y\rangle \stackrel{\mathrm{def}}{=} \mathrm{Cov}(X,Y)$ (la notazione è giustificata dal fatto che la covarianza soddisfa gli assiomi di un prodotto scalare). Allora per la disuguaglianza di Cauchy-Schwarz,

$$|\langle X,Y\rangle| \leq \langle X,X\rangle^{1/2}\langle Y,Y\rangle^{1/2}$$

e quindi

$$|\mathrm{Cov}(X,Y)| \leq \mathrm{Var}(X)^{1/2}\mathrm{Var}(Y)^{1/2}$$

ovvero

$$-\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} \leq \mathrm{Cov}(X,Y) \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

□

Definizione 4.1.4. Il *coefficiente di correlazione empirico* tra due sequenze $x = (x_1, \ldots, x_m)$ e $y = (y_1, \ldots, y_m)$ è dato da

$$r_{x,y} \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^m (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^m (x_i-\bar{x})^2 \sum_{i=1}^m (y_i-\bar{y})^2}}$$

dove $\bar{x} = (1/m)\sum_{i=1}^m x_i$, $\bar{y} \stackrel{\mathrm{def}}{=} (1/m)\sum_{i=1}^m y_i$.

Il coefficiente di correlazione empirico tra x ed y equivale al coefficiente di correlazione tra le due variabili aleatorie X ed Y definite da $\Pr[X = x_i] = 1/m$ e $\Pr[Y = y_i] = 1/m$ per ogni $i = 1, 2, \ldots, m$. Di conseguenza si ha $-1 \leq r_{x,y} \leq 1$.

4.2 Matrici di covarianza

La *matrice di covarianza* associata a d variabili aleatorie reali $X_1, \ldots, X_d$ è la matrice simmetrica $\Sigma \in \mathbb{R}^{n \times n}$ definita da

$$\Sigma_{ij} \stackrel{\text{def}}{=} \mathrm{Cov}(X_i, X_j).$$

Teorema 4.2.1. *Ogni matrice di covarianza Σ è semidefinita positiva.*

Dimostrazione. Come già osservato, la covarianza tra due variabili aleatorie soddisfa le proprietà di un prodotto scalare, dunque ogni matrice di covarianza può essere scritta nella forma (di Gram)

$$\Sigma_{ij} = \langle X_i, X_j \rangle$$

e dunque per ogni vettore $u \in \mathbb{R}^d$,

$$u^\top \Sigma u = \sum_{i,j} \langle X_i, X_j \rangle u_i u_j = \sum_{i,j} \langle u_i X_i, u_j X_j \rangle = \langle \sum_i u_i X_i, \sum_j u_j X_j \rangle = \left\| \sum_i u_i X_i \right\|^2 \geq 0.$$

$\square$