1. REGRESSIONE LINEARE: LA RETTA DEI MINIMI QUADRATI

Siano assegnate n coppie di dati (punti di $\mathbb{R}^2$)

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

e si consideri il problema di determinare l'equazione di una retta

$$y = ax + b$$

in corrispondenza della quale risulti minima la quantità

$$\epsilon(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$\epsilon(a, b)$ è una funzione convessa della variabili (a, b) che tende a $+\infty$ per $(a, b) \to \infty$ e pertanto ammette uno ed un solo punto di minimo assoluto che si può trovare annullando $\nabla\epsilon$.

Per risolvere il problema dovremo pertanto risolvere il sistema definito dalle equazioni

$$\begin{cases} \frac{\partial \epsilon}{\partial a} = \sum_{i=1}^{n} -2(y_i - ax_i - b)x_i = 0 \\ \frac{\partial \epsilon}{\partial b} = \sum_{i=1}^{n} -2(y_i - ax_i - b) = 0 \end{cases}$$

Ne viene che

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} 1 = 0 \end{cases}$$

ovvero

$$\boxed{\begin{cases} \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb \end{cases}} \tag{1.1}$$

Dalla seconda delle 1.1 si può vedere che

$$nb = \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i$$

ed anche

$$\boxed{b = \frac{\sum_{i=1}^{n} y_i}{n} - a\frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - a\bar{x}}$$

dove

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad , \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

indicano la media dei valori x_i ed y_i, rispettivamente.

Dalla prima delle 1.1 si può invece ottenere che

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \\ = a \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i \left(\frac{\sum_{i=1}^{n} y_i}{n} - a\frac{\sum_{i=1}^{n} x_i}{n} \right)$$

e

$$\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = a \left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right)$$

$$n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = a \left(n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right)$$

ed infine

$$\boxed{a = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

Inoltre

$$nb = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \left(\frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \right) = \\ = \frac{1}{n \left(\sum_{i=1}^{n} x_i\right)^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \left[n \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} x_i \right)^2 \sum_{i=1}^{n} y_i \right. \\ \left. -n \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i + \left(\sum_{i=1}^{n} x_i \right)^2 \sum_{i=1}^{n} y_i \right)$$

e se ne conclude che

$$\boxed{b = \frac{n \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

Ora, tenendo conto che

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}\bar{x}y_i-\sum_{i=1}^{n}x_i\bar{y}+\sum_{i=1}^{n}\bar{x}\bar{y}=$$
$$=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y}=$$
$$=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}=\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}$$

e che

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}\bar{x}x_i+\sum_{i=1}^{n}\bar{x}^2=$$
$$=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

si ricava che

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=$$
$$=\frac{1}{n}\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=$$
$$=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-n^2\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2}=a$$

Pertanto possiamo esprimere a e b mediante le seguenti formule

$$\boxed{\begin{cases}a=\dfrac{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})^2}\\ \bar{y}=a\bar{x}+b\end{cases}} \tag{1.2}$$

La prima delle due uguaglianze premette di concludere che a è invariante rispetto alla traslazione degli assi: cioè usando $x-x_0$ ed $y-y_0$ in luogo di x ed y il valore di a non cambia.

La stessa trasformazione cambia invece il valore di b, come si vede dalla seconda uguaglianza. Dalla medesima si vede anche che la retta di regressione passa per il punto di coordinate $(\bar{x},\bar{y})$ che è il baricentro dei dati.

Possiamo anche osservare che, a meno di operare una traslazione dei dati riportando l'origine degli assi nel baricentro $(\bar{x},\bar{y})$,si può sup-

porre che

$$\boxed{\begin{cases} a = \dfrac{\displaystyle\sum_{i=1}^{n} x_i y_i}{\displaystyle\sum_{i=1}^{n} x_i^2} \\ b = 0 \end{cases}} \tag{1.3}$$

Ora, siano

- s_x^2 la varianza dei dati x_i

$$\boxed{s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

- s_y^2 la varianza dei dati y_i

$$\boxed{s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

- s_{xy} la covarianza dei dati (x_i, y_i)

$$\boxed{s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}}$$

Possiamo scrivere la retta di regressione nella forma

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

e se invertiamo il ruolo di x e di y l'equazione diventa

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

Possiamo misurare la correlazione tra i dati utilizzando il coefficiente definito da

$$\boxed{r = \frac{s_{xy}}{s_x s_y}} \tag{1.4}$$

mediante il quale le equazioni delle due rette prima introdotte diventano

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

e

$$r\frac{y - \bar{y}}{s_y} = \frac{x - \bar{x}}{s_x}$$

Chiaramente le due rette coincidono soltanto nel caso in cui

$$r^2 = 1 \qquad \text{cioè} \qquad r = \pm 1$$

e il fatto che questo accada è indice della correlazione dei dati cioè del fatto che i dati si trovano su una retta.

È ragionevole quindi stimare la maggiore o minore correlazione tra i dati confrontando r^2 con 1: più r^2 è vicino ad 1 e più i dati sono da considerarsi linearmente correlati.

Possiamo inoltre misurare la dispersione dei dati attorno alla retta di regressione mediante la

$$S_{y,x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^s)^2}{n}}$$

dove

$$y_i^s = ax_i + b$$

Pertanto

$$S_{y,x}^2 = \frac{\sum_{i=1}^{n}(y_i - ax_i - b)^2}{n}$$

e

$$\frac{\sum_{i=1}^{n}(y_i - ax_i - b)^2}{n} = \sum_{i=1}^{n}(y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i) =$$

$$= \sum_{i=1}^{n} y_i^2 + a^2\sum_{i=1}^{n} x_i^2 + nb^2 - 2a\sum_{i=1}^{n} x_iy_i - 2b\sum_{i=1}^{n} y_i + 2ab\sum_{i=1}^{n} x_i) =$$

e per le 1.1

$$= \sum_{i=1}^{n} y_i^2 + a^2\left[\frac{1}{a}\left(\sum_{i=1}^{n} x_iy_i - b\sum_{i=1}^{n} x_i\right)\right] + nb^2 -$$

$$- 2a\sum_{i=1}^{n} x_iy_i - 2b\sum_{i=1}^{n} y_i + 2ab\left[\frac{1}{a}\left(\sum_{i=1}^{n} y_i - nb\right)\right] =$$

$$= \sum_{i=1}^{n} y_i^2 + a\sum_{i=1}^{n} x_iy_i - ab\sum_{i=1}^{n} x_i + nb^2 -$$

$$- 2a\sum_{i=1}^{n} x_iy_i - 2b\sum_{i=1}^{n} y_i + 2b\sum_{i=1}^{n} y_i - 2nb^2 =$$

$$= \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_iy_i - ab\sum_{i=1}^{n} x_i - nb^2 =$$

$$= \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_iy_i - b\left(a\sum_{i=1}^{n} x_i + nb\right) =$$

e ancora per le 1.1

$$= \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_iy_i - b\sum_{i=1}^{n} y_i$$

Quindi

$$S_{y,x}^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_iy_i - b\sum_{i=1}^{n} y_i}{n}$$

Si può d'altro canto verificare che

$$\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(y_i - \bar{y})^2 - a\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

infatti

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \bar{y})^2 - a\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \\
&= \sum_{i=1}^{n} y_i^2 - 2\bar{y}\sum_{i=1}^{n} y_i + n\bar{y}^2 - a\sum_{i=1}^{n} x_i y_i + a\sum_{i=1}^{n} x_i\bar{y} + \\
&\quad + a\sum_{i=1}^{n}\bar{x}y_i - a\sum_{i=1}^{n}\bar{x}\bar{y} = \\
&= \sum_{i=1}^{n} y_i^2 - 2n\bar{y}^2 + n\bar{y}^2 - a\sum_{i=1}^{n} x_i y_i + an\bar{x}\bar{y} + an\bar{x}\bar{y} - an\bar{x}\bar{y} = \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - a\sum_{i=1}^{n} x_i y_i + an\bar{x}\bar{y} = \\
&= \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i + n\bar{y}(\bar{y} - a\bar{x}) = \\
&= \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i + n\bar{y}b = \\
&= \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i + b\sum_{i=1}^{n} y_i
\end{aligned}$$

Le precedenti considerazioni permettono quindi di affermare che

$$\begin{aligned}
S_{y,x}^2 &= \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} y_i}{n} = \\
&= \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - a\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = s_y^2 - a s_{xy}
\end{aligned}$$

e dal momento che $a = \frac{s_{xy}}{s_x^2}$

$$\begin{aligned}
&= s_y^2\left(1 - a\frac{s_{xy}}{s_y^2}\right) = s_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right) = \\
&s_y^2\left(1 - r^2\right)
\end{aligned}$$

Ne viene quindi che

$$\boxed{\frac{S_{y,x}^2}{s_y^2} = (1 - r^2)}$$

e

$$\boxed{r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^s)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

D'altro canto

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-y_i^s+y_i^s-\bar{y})^2=$$
$$=\sum_{i=1}^{n}(y_i-y_i^s)^2+\sum_{i=1}^{n}(y_i^s-\bar{y})^2+2\sum_{i=1}^{n}(y_i-y_i^s)(y_i^s-\bar{y})$$

Poichè valgono le equazioni normali 1.1 che definiscono a e b

$$\sum_{i=1}^{n}(y_i-y_i^s)(y_i^s-\bar{y})=\sum_{i=1}^{n}(y_i-ax_i-b)(ax_i+b-\bar{y})=$$
$$=(b-\bar{y})\sum_{i=1}^{n}(y_i-ax_i-b)+a\sum_{i=1}^{n}x_i(y_i-ax_i-b)$$
$$=(b-\bar{y})\left(\sum_{i=1}^{n}y_i-a\sum_{i=1}^{n}x_i-nb\right)+a\left[\sum_{i=1}^{n}x_iy_i-a\sum_{i=1}^{n}x_i^2-b\sum_{i=1}^{n}x_i\right]=0$$

avremo

$$r^2=1-\frac{\sum_{i=1}^{n}(y_i-y_i^s)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-y_i^s)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=$$
$$=\frac{\sum_{i=1}^{n}(y_i^s-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\text{Variazione spiegata}}{\text{Variazione totale}}$$

Possiamo anche calcolare, dalla 1.4, che

$$r=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left(n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\right)\left(n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2\right)}}=$$
$$=\frac{\overline{xy}-\bar{x}\bar{y}}{\sqrt{\left(\overline{x^2}-\bar{x}^2\right)\left(\overline{y^2}-\bar{y}^2\right)}}=$$
$$=\frac{\sum_{i=1}^{n}x_iy_i}{\sqrt{(\sum_{i=1}^{n}{x_i}^2)(\sum_{i=1}^{n}{y_i}^2)}}$$

2. ANALISI DEI COMPONENTI PRINCIPALI.

2.1 Forme quadratiche ed autovalori.

Sia A una matrice $n \times n$ e consideriamo la funzione

$$f(u) = \langle Au, u \rangle \qquad , \quad u \in \mathbb{R}^n$$

f si chiama forma quadratica su $\mathbb{R}^n$ e si può vedere che è sempre possibile supporre che la matrice A che la individua sia simmetrica.

Infatti, ad esempio per $n = 2$, se

$$A_1 = \begin{pmatrix} a & d_1 \\ d_2 & c \end{pmatrix}$$

ed $u = \begin{pmatrix} x \\ y \end{pmatrix}$ avremo che

$$f(u) = \langle A_1 u, u \rangle = ax^2 + d_1 xy + d_2 yx + cy^2 = ax^2 + (d_1 + d_2)xy + cy^2 = \\ = ax^2 + 2bxy + cy^2 = \langle Au, u \rangle$$

per $b = (d_1 + d_2)/2$, da cui

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

D'altro canto, se A è una matrice simmetrica possiamo verificare che

$$\langle Au, v \rangle = \langle u, Av \rangle$$

ed inoltre

$$f(u+h) = \langle A(u+h), (u+h) \rangle = \langle Au, u \rangle + 2\langle Au, h \rangle - \langle Ah, h \rangle$$

per cui

$$f(u+h) - f(u) = 2\langle Au, h \rangle + \langle Ah, h \rangle = 2\langle Au, u \rangle + \|h\| \omega(h)$$

con ω funzione infinitesima per $h \to 0$ (si ricordi che $|\langle Ah, h\rangle| \leq \|Ah\|\|h\|$ ed $Ah \to 0$ se $h \to 0$).

Dalla definizione di differenziale si ottiene allora che f è differenziabile e che

$$\nabla f(u) = 2Au$$

Come caso particolare, se $A = I$ (la matrice identica) si ha

$$g(u) = \langle u, u\rangle = \|u\|^2 \quad , \quad \nabla g(u) = 2u$$

Consideriamo ora il problema di trovare

$$\max_{g(u)-1=0} f(u) = \max_{\|u\|^2=1} \langle Au, u\rangle$$

Per il teorema di Weierstraß, dal momento che f è continua e che $\|u\|^2 = 1$ definisce la superficie della sfera di centro l'origine e raggio 1, che è chiusa e limitata, possiamo affermare che il massimo esiste. Sia u_1 il punto in cui tale massimo è assunto

$$\langle Au_1, u_1\rangle = \max_{\|u\|^2=-1} \langle Au, u\rangle$$
$$\|u_1\|^2 = 1$$

D'altro canto, il teorema dei moltiplicatori di Lagrange consente di affermare che esiste λ_1 tale che

$$\nabla f(u_1) = \lambda_1 \nabla g(u_1)$$

per cui deve essere

$$Au_1 = \lambda_1 u_1$$

Dal momento che la precedente equazione è soddisfatta

- λ_1 è un autovalore di A
- u_1 è un autovettore di A corrispondente all'autovalore λ_1

Possiamo inoltre osservare che

$$\max_{\|u\|^2=1} \langle Au, u\rangle = \langle Au_1, u_1\rangle = \langle \lambda_1 u_1, u_1\rangle = \lambda_1 \|u_1\|^2 = \lambda_1$$

per cui λ_1 è il valore massimo assunto da $\langle Au, u\rangle$ sulla sfera $\|u\|^2 = 1$

Consideriamo ora lo spazio vettoriale V_1 generato da u_1

$$V_1 = \{\lambda u_1 \ : \ \lambda \in \mathbb{R}\}$$

e lo spazio $V_1^\perp$ ortogonale a V_1

$$V_1^\perp = \{v \in \mathbb{R}^n \ : \ \langle v, u_1\rangle = 0 \ \} = \{v \in \mathbb{R}^n \ : \ h(v) = 0 \ \}$$

con $h(v) = \langle v, u_1 \rangle$.

Consideriamo ora il problema di trovare

$$\max_{\substack{g(u)-1=0 \\ h(u)=0}} f(u) = \max_{\substack{\|u\|^2=1 \\ \langle u,u_1\rangle=0}} \langle Au, u \rangle$$

possiamo anche qui applicare il metodo dei moltiplicatori di Lagrange ed affermare che che esistono $u_2 \in V_1^{\perp}$ e λ_2, μ_2 tali che

$$\nabla f(u_2) = \lambda_2 \nabla g(u_2) + \mu_2 \nabla h(u_2)$$

ma $\nabla h(u) = u_1$ in quanto h è lineare, e otteniamo

$$2Au_2 = 2\lambda_2 u_2 + \mu_2 u_1$$

da cui

$$2(A - \lambda_2)u_2 = \mu_2 u_1$$

Moltiplicando per u_1^t otteniamo

$$0 = 2(A - \lambda_2)u_2 u_1^t = \mu_2 \|u_1\| = \mu_1$$

da cui $\mu_1 = 0$ e

$$(A - \lambda_2)u_2 = 0$$

Ne deduciamo che λ_2 è autovalore di A e u_2 è l'autovettore corrispondente e che $\lambda_1 \geq \lambda_2$ in quanto $\{v : g(u) - 1 = 0, h(u) = 0\} \subset \{v : g(u) - 1 = 0\}$.

Possiamo iterare i procedimento

Consideriamo ora lo spazio V_2 generato da u_1, u_2

$$V_2 = \{\lambda u_1 + \mu_2 \ : \ \lambda, \mu \in \mathbb{R}\}$$

e lo spazio $V_1^{\perp}$ ortogonale a V_1

$$V_2^{\perp} = \{v \in \mathbb{R}^n \ : \ l(v) = 0 \ \}$$

con $h(v) = Lv$, $L = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$.

Consideriamo il problema di trovare

$$\max_{\substack{g(u)-1=0 \\ A(u)=0}} f(u) = \max_{\substack{\|u\|^2=1 \\ Au=0}} \langle Au, u \rangle$$

possiamo ancora applicare il metodo dei moltiplicatori di Lagrange ed affermare che che esistono $u_3 \in V_2^{\perp}$ e $\lambda_3, (\mu_3, \eta_3)$ tali che

$$\nabla f(u_3) = \lambda_3 \nabla g(u_3) + (\mu_3, \eta_3) \nabla l(u_3)$$

ma $\nabla l(u) = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ in quanto l è lineare, e otteniamo

$$2Au_3 = 2\lambda_3 u_3 + (\mu_3, \eta_3) \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

da cui

$$2(A - \lambda_3)u_3 = \mu_3 u_1 + \eta_3 u_2$$

Moltiplicando per $(\mu_3 u_1 + \eta_3 u_2)^t$ otteniamo

$$0 = 2(A - \lambda_3)u_3(\mu_3 u_1 + \eta_3 u_2)^t = \|\mu_3 u_1 + \eta_3 u_2\|^2$$

da cui $\mu_3 u_1 + \eta_3 u_2 = 0$ e

$$(A - \lambda_3)u_3 = 0$$

Ne deduciamo che λ_3 è autovalore di A e u_3 è l'autovettore corrispondente e che $\lambda_1 \geq \lambda_2 \geq \lambda_3$ in quanto

$$\{v : g(u) - 1 = 0, l(u) = 0\} \subset \{v : g(u) - 1 = 0, h(u) = 0\} \subset \{v : g(u) - 1 = 0\}$$

Chiaramente si può ripetere quanto fatto fino a trovare:

- n autovalori $\lambda_1, \lambda_2, ..., \lambda_n$ decrescenti in valore
- n autovettori $u_1, u_2, ..., u_n$, uno per ogni autovalore, che risultano ortogonali tra loro e di norma unitaria.

Gli autovettori $u_1, u_2, ..., u_n$ formano quindi una base ortonormale con la caratteristica che lungo il primo asse si trova il punto di massimo della forma quadratica $\langle Au, u \rangle$ sulla sfera unitaria in $\mathbb{R}^2$, lungo il secondo asse si trova il massimo della forma quadratica $\langle Au, u \rangle$ sulla sfera unitaria in $V_1^{\perp}$ e così via fino all'$n-$esimo asse.

Infatti, sia

$$R = \begin{pmatrix} u_1^1 & u_2^1 & \dots & u_n^1 \\ u_1^2 & u_2^2 & \dots & u_n^2 \\ \dots & \dots & \ddots & \dots \\ u_1^n & u_2^n & \dots & u_n^3 \end{pmatrix} = \begin{pmatrix} u_1 & u_2 & \dots & u_n \end{pmatrix}$$

la matrice che ha per colonne i vettori u_i; R è una matrice ortonormale e rappresenta una rotazione in $\mathbb{R}^n$.

2.2 *Analisi delle componenti principali. PCA.*

Vediamo ora di formalizzare quanto abbiamo potuto vedere nell'esempio svolto.

Sia A una matrice $n \times p$ che raccoglie n osservazioni relative a p variabili. Per facilitare la comprensione supporremo $p = 3$, osservando esplicitamente che la presenza di più di tre variabili comporta soltanto un aggravio delle notazioni.

$$A = \begin{pmatrix} x_1^1 & x_2^1 & x_p^1 \\ x_1^2 & x_2^2 & x_p^2 \\ x_1^3 & x_2^3 & x_p^3 \\ \dots & \dots & \dots \\ x_1^n & x_2^n & x_p^n \end{pmatrix}$$

e consideriamo la matrice di covarianza dei dati che è definita da

$$C = A^t A$$

e risulta definita da

$$C = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 \end{pmatrix}$$

dove

$$\sigma_{ij}^2 = \sum_k (x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j)$$

essendo chiaramente $\bar{x}_i$ la media di x_i, $\sigma_{i,i}$ la varianza di x_i e $\sigma_{i,j}$ la covarianza di x_i ed x_j.

La matrice C risulta una matrice simmetrica ed è noto che esiste una matrice diagonale

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

ed una matrice ortonormale

$$R = \begin{pmatrix} u_1^1 & u_2^1 & u_3^1 \\ u_1^2 & u_2^2 & u_3^2 \\ u_1^3 & u_2^3 & u_3^3 \end{pmatrix} = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix}$$

tale che

$$R^t C R = D$$

La matrice D presenta sulla diagonale principale gli autovalori λ_i reali e non negativi di C, mentre le colonne di R sono costituite dagli autovettori $u_i = (u_i^1, u_i^2, u_i^3)$ corrispondenti.

Più esplicitamente si ha

$$\begin{pmatrix} u_1^1 & u_1^2 & u_1^3 \\ u_2^1 & u_2^2 & u_2^3 \\ u_3^1 & u_3^2 & u_3^3 \end{pmatrix} \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 \end{pmatrix} \begin{pmatrix} u_1^1 & u_2^1 & u_3^1 \\ u_1^2 & u_2^2 & u_3^2 \\ u_1^3 & u_2^3 & u_3^3 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

e quindi svolgendo i calcoli, per l'ortogonalità delle colonne di R (e delle righe di R^t), si ha

$$(AR)^t AR =$$
$$= R^t C R = \begin{pmatrix} < Cu_1, u_1 > & 0 & 0 \\ 0 & < Cu_2, u_2 > & 0 \\ 0 & 0 & < Cu_3, u_3 > \end{pmatrix} =$$
$$= \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

Pertanto la forma quadratica associata alla matrice di covarianza $(AR)^t AR$ relativa ai dati contenuti in A ruotati mediante R è data da

$$\varphi(a, b, c) = \lambda_1 a^2 + \lambda_2 b^2 + \lambda_3 c^2$$

e si vede che , posto

$$\Lambda = \max\{\lambda_1, \lambda_2, \lambda_3\} \quad \text{e} \quad \lambda = \min\{\lambda_1, \lambda_2, \lambda_3\}$$

si ha

$$\lambda \leq \lambda_1 a^2 + \lambda_2 b^2 + \lambda_3 c^2 \leq \Lambda \qquad , \qquad \forall (a, b, c) \in \mathbb{R}^3 \ , \ a^2 + b^2 + c^2 = 1$$

Ciò assicura che la forma quadratica associata alla matrice di covarianza relativa ai dati ruotati risulta massima in corrispondenza della direzione individuata dall'autovettore associato al massimo autovalore.

L'uguaglianza

$$(AR)^t AR = R^t A^T AR = R^t C R = D$$

permette anche di concludere che, se consideriamo una generica riga della matrice A $a = \begin{pmatrix} x & y & z \end{pmatrix}$ otteniamo

$$AR = \begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} = \begin{pmatrix} < a, u_1 > & < a, u_2 > & < a, u_3 > \end{pmatrix}$$

per cui

$$(AR)^t AR = \begin{pmatrix} < a, u_1 >^2 & 0 & 0 \\ 0 & < a, u_2 >^2 & 0 \\ 0 & 0 & < a, u_2 >^2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

Ne segue che, se λ_i è trascurabile

$$< a, u_i >= 0$$

Ciò permette di determinare una relazione lineare tra le variabili che sono riportate nelle colonne di A

2.3 *L'applicazione all'analisi delle componenti principali.*

Torniamo ora ai nostri dati $P_k = (u_k^1, u_k^2, ..., u_k^n) \in \mathbb{R}^n$ e cerchiamo di individuare una combinazione lineare delle componenti

$$Q_k = \alpha_1 u_k^1 + \alpha_2 u_k^2 + ... + \alpha_n u_k^n$$

in modo che la varianza di Q_k sia massima (massima significatività della variabile).

Definiamo

$$v_i = \mathrm{Var}(u_k^i)$$
$$c_{ij} = \mathrm{Cov}(u_k^i, u_k^j)$$

la varianza delle singole componenti e la covarianza delle componenti a due a due e sia R la **matrice di covarianza** dei dati definita mediante la

$$R = \begin{pmatrix} v1 & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & v2 & c_{23} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & v_n \end{pmatrix}$$

Possiamo allora verificare che

$$\mathrm{Var}(Q_k) = \langle Ra, a \rangle$$

dove

$$a = (\alpha_1, \alpha_2, ..., \alpha_n)$$

$\mathrm{Var}(Q_k)$ è quindi una forma quadratica cui possiamo applicare il metodo visto nella precedente sezione e mediante tale metodo possiamo individuare in ordine decrescente di significatività le componenti dei dati.

2.3.1 *Un esempio*

Per illustrare gli effetti del metodo consideriamo i punti del grafico di $\sin(t)$ nell'intervallo $[0, 2\pi]$; suddividiamo l'intervallo in 199 parti uguali, in modo da individuare 200 punti in $[0, 2\pi]$ e calcoliamo i valori assunti da $\sin(t)$ in tali punti.

Le seguenti istruzioni di Matlab producono come risultato un vettore t che contiene i 200 punti in $[0, 2\pi]$, ed i vettori x ed y che contengono i valori assunti da $\sin(6t)$ nei punti pari e dispari rispettivamente

```
clear all;
step=2*pi/199;
t=0:step:2*pi;
```

```
xx=0:2*step:  2*pi;
yy=step:2*step:  2*pi;
x=sin(6*xx);
y=sin(6*yy);
```

Possiamo esaminare nel piano la distribuzione dei punti di coordinate $(x(k), y(k))$ con $k = 1, .., 100$ mediante le seguenti istruzioni

```
figure(1)
plot(x,y,'+');
axis([-2 2 -2 2]);
```

che producono la seguente figura 2.1

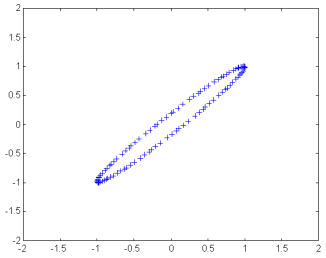

Figure 2.1: .

Le istruzioni

```
R=cov(x,y);
[V,D]=eig(R)
```

calcolano la matrice di covarianza R e le matrici V e D , dove V è la matrice le cui colonne sono gli autovettori di R e D è una matrice diagonale con gli autovalori sulla diagonale principale; in altre parole V è la matrice tale che

$$VR = RD \quad \text{cioè tale che} \quad V^{-1}RV = D$$

La matrice V pertanto è la matrice di passaggio dal sistema di coordinate originale a quello individuato dagli autovalori di R.

Poichè ci interessa tener conto della componente relativa al massimo autovalore, di assicuriamo anche che l'autovalore più grande sia in posizione $(1, 1)$ mediante le istruzioni

```
if D(1,1)>D(2,2)
vv=V(:,1);
V(:,1)=V(:,2);
V(:,2)=vv;
end
```

La matrice V, quindi, può essere usata per effettuare un cambio di base che metta in evidenza le componenti principali.

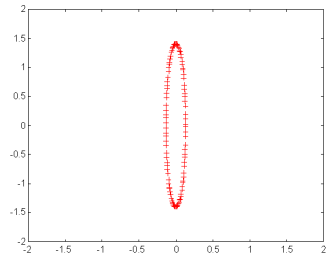

Figure 2.2: .

Le seguenti istruzioni

```
tr=[x(:),y(:)]*V;
tr1=tr(:,1);
tr2=tr(:,2);
figure(2)
plot(tr1,tr2,'r+')
axis([-2 2 -2 2]);
```

Calcolano i trasformati $tr1$, $tr2$ dei punti (x, y) e li mostrano rispetto ad una coppia di assi ortogonali coincidenti con gli autovettori di R (si veda la figura 2.2).

Le successive istruzioni:

```
rtr=[tr1(:),tr2(:)]*inv(V);
rtr1=rtr(:,1);
```

```
rtr2=rtr(:,2);
figure(3)
plot(rtr1,rtr2,'g+')
axis([-2 2 -2 2]);
```

mostrano come applicando la trasformazione inversa i dati possano essere recuperati (si veda la figura 2.3).

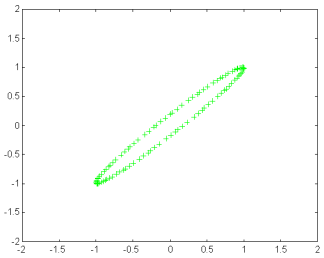

Figure 2.3: .

Ora, possiamo osservare che la variazione della seconda componente dei dati trasformati è trascurabile rispetto alla prima, per cui, se la trascuriamo e applichiamo la trasformazione inversa ai dati privati di tale componente, otteniamo nuovi punti che differiscono di poco da quelli originali; possiamo operare usando le seguenti istruzioni:

```
nu=zeros(size(tr2));
ntr=[nu(:),tr2(:)]*inv(V);
ntr1=ntr(:,1);
ntr2=ntr(:,2);
figure(4)
plot(ntr1,ntr2,'rx',rtr1,rtr2,'g+')
axis([-2 2 -2 2]);
```

che forniscono anche una immagine dei nuovi punti come indicato in figura 2.3 ed un confronto con i punti originali 2.4.

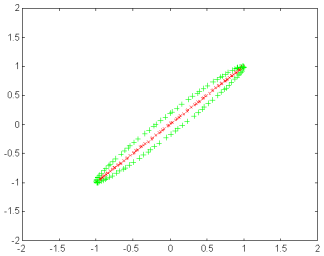

Figure 2.4: .

È interessante ora osservare come dai punti originali e da quelli privati della componente meno significativa si possa ricostruire la funzione $\sin(6t)$

Le seguenti istruzioni

```
z=zeros(1,200);
zt=zeros(1,200);
for k=0:99
zt(2*k+1)=x(k+1);
end
for k=1:100
zt(2*k)=y(k);
end
for k=0:99
z(2*k+1)=ntr2(k+1);
end
for k=1:100
z(2*k)=ntr1(k);
end
figure(5)
plot(t,z)
figure(6)
plot(t,zt)
```

producono i due grafici riportati in figura **??**, il primo dei quali riporta la funzione $\sin(t)$ ricostruita congiungendo con segmenti di retta

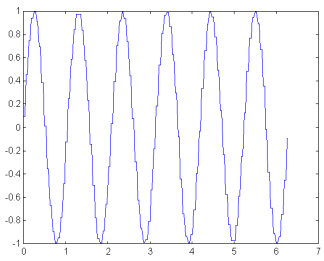

Figure 2.5: .

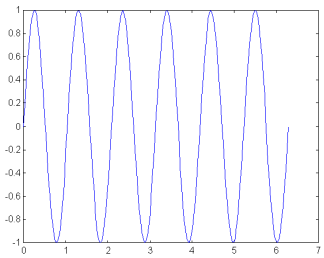

Figure 2.6: .

i 200 punti originali, mentre la seconda riporta il grafico di $\sin(6t)$ ricostruito a partire dai punti ottenuti applicando la trasformazione inversa ai punti prima trasformati e poi privati della seconda componente.

Come si vede è evidente che la componente trascurata non ha peggiorato di molto il grafico, mentre la quantità di dati necessari a ricostruire l'immagine si è dimezzata. Questo indica come può essere sviluppato un procedimento che consenta di immagazzinare dati (i punti del grafico della funzione) utilizzando al meglio le informazioni che contengono.