STATISTICA MULTIVARIATA

Capitoli

Varia

La matrice dei dati

Regressione e correlazione

Il teorema spettrale

Analisi delle componenti principali

Rappresentazioni grafiche

Regressione multivariata

Ottimizzazione genetica

Raggruppamento automatico

Difficoltà in alta dimensione

R

Python

I. LA RETTA DI REGRESSIONE

Il principio di dualità

Assumiamo che i valori di due variabili numeriche (ad es. le concentrazioni di due aminoacidi nel sangue) siano stati misurati per n oggetti o individui (ad es. pazienti); otteniamo così n punti $(x^1, y^1), \dots, (x^n, y^n)$ nel piano $\mathbb{R}_2$ che possono essere rappresentati da una matrice

$$\begin{pmatrix} x^1 & y^1 \\ x^2 & y^2 \\ \vdots & \vdots \\ x^n & y^n \end{pmatrix}$$

a 2 colonne ed n righe. Questa matrice si chiama la *matrice dei dati*.

Le righe (x^i, y^i) forniscono da sole tutta l'informazione contenuta nella matrice, così come le colonne. Ciononostante guardando solo le righe o solo le colonne, in un certo senso si vede solo la metà di questa informazione; l'altra metà è nascosta, difficile da comprendere. Solo lavorando contemporaneamente con righe e colonne tutta l'informazione appare sempre chiaramente davanti ai nostri occhi.

Ciò è tipico per *situazioni di dualità*, in cui due aspetti di uno stesso oggetto o di una stessa struttura si determinano reciprocamente in modo (più o meno) completo e in cui quindi ogni enunciato su uno dei due aspetti implica un enunciato anche sull'altro aspetto, e dove ciononostante spesso questi due enunciati devono essere formulati o dimostrati in modo apparentemente molto diverso.

Può così accadere che in uno dei due aspetti un enunciato o un algoritmo si presentino in veste molto semplice e diventino molto più difficili quando vengono tradotti nell'altro aspetto. È quindi spesso preferibile tener presente i due aspetti contemporaneamente invece di cercare di ridurre l'uno all'altro: per definizione ciò sarebbe possibile, ma a spese della comprensione.

Uno dei più noti esempi di dualità è l'analisi di Fourier; il buon analista di Fourier ha sempre davanti agli occhi entrambi gli aspetti della dualità e non preferisce nessuno dei due.

In questo spirito introduciamo adesso, partendo dalla nostra matrice di dati, le colonne

$$x = \begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix} \qquad y = \begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix}$$

come nuovi oggetti. x e y come punti sono elementi di un $\mathbb{R}^n$ a dimensione molto alta (ad esempio $n = 50000$ in uno screening di 50000 neonati); la loro geometria implica e chiarisce talvolta circostanze per i dati in $\mathbb{R}_2$ che sarebbe difficile individuare direttamente nel piano dei dati.

Dipendenza funzionale

In matematica il concetto di funzione è definito in modo molto generale. Se in una tabella come

$$\begin{pmatrix} 3 & 2 \\ 5 & 1 \\ 2 & 8 \\ 1 & 0 \\ 6 & 0 \\ 8 & 2 \end{pmatrix}$$

gli elementi della prima colonna sono tutti distinti, ciò è sufficiente per poter considerare la seconda colonna come funzione della prima: definiamo una funzione $f : \{3, 5, 2, 1, 6, 8\} \longrightarrow \{0, 1, 2, 8\}$ semplicemente ponendo $f(3) = 2$, $f(5) = 1$, $f(2) = 8$, $f(1) = f(6) = 0$, $f(8) = 2$. In questo senso quindi la seconda colonna dipende in modo funzionale dalla prima, benché si possa difficilmente affermare l'esistenza di qualche legame statistico o addirittura causale tra le due variabili. Solo quando la funzione appartiene a una classe determinata e possibilmente semplice di funzioni (lineari, quadratiche, logaritmiche, monotone, sigmoidali, sinusoidali) si può cercare di associare a una tale relazione un significato statistico.

Quindi anche in una rappresentazione grafica dei dati nel piano, in cui i valori x^i sono tutti distinti, ciò da solo ci permette di considerare i valori y^i come funzione degli x^i nel senso della matematica.

La matrice dei dati

Definizione 1.1. Sia $X \in \mathbb{R}^n_m$ con $n \geq 2$. Scriviamo X nella forma

$$X = \begin{pmatrix} X^1_1 & \dots & X^1_m \\ \vdots & & \vdots \\ X^n_1 & \dots & X^n_m \end{pmatrix}$$

La j-esima colonna di X è denotata con X_j. Abbiamo quindi

$$X_j = \begin{pmatrix} X^1_j \\ \vdots \\ X^n_j \end{pmatrix} \quad \text{ed} \quad X = (X_1, \dots, X_m).$$

La i-esima riga di X è invece $X^i := (X^i_1, \dots, X^i_m)$.

Nota 1.2. Nel caso $m = 2$, che considereremo nei primi capitoli, scriveremo spesso $X_1 = x = (x^1, \dots, x^n)^t$, $X_2 = y = (y^1, \dots, y^n)^t$. In questo caso $X = (x, y)$.

Vettori diagonali e operatori di ripetizione

Definizione 1.3. Un vettore di $\mathbb{R}^n$ o di $\mathbb{R}^m$ si chiama *diagonale*, se tutti i suoi coefficienti sono uguali.

Definizione 1.4. Sia $f \in \mathbb{R}_m$ un vettore riga. Allora con

$$f^\diamond := \begin{pmatrix} f \\ \vdots \\ f \end{pmatrix} \in \mathbb{R}^n_m$$

denotiamo la matrice che si ottiene ripetendo n volte la riga f. Similmente per un vettore colonna $v \in \mathbb{R}^n$ definiamo il vettore $v_\diamond := (v, \dots, v) \in \mathbb{R}^n_m$ come la matrice che si ottiene ripetendo m volte la colonna v.

Un numero $\lambda \in \mathbb{R}$ può essere considerato sia come vettore riga che come vettore colonna, perciò sono definiti i vettori diagonali

$$\lambda^\diamond = \begin{pmatrix} \lambda \\ \vdots \\ \lambda \end{pmatrix} \in \mathbb{R}^n \text{ e } \lambda_\diamond = (\lambda, \dots, \lambda) \in \mathbb{R}_m$$

I vettori $1^\diamond$ e $1_\diamond$, che chiamiamo *vettori diagonali unitari* di $\mathbb{R}^n$ e $\mathbb{R}_m$, sono molto utili nella statistica geometrica. Il simbolo $\diamond$ è pronunciato *"diagonale"*; gli operatori $^\diamond$ e $_\diamond$ si chiamano *operatori di ripetizione*.

La retta $\mathbb{R}^\diamond := \mathbb{R}1^\diamond$ si chiama la *retta diagonale* di $\mathbb{R}^n$; similmente è definita la retta diagonale $\mathbb{R}_\diamond := \mathbb{R}1_\diamond$ di $\mathbb{R}_m$.

Osservazione 1.5. $|1^\diamond| = \sqrt{n}$.

La media

Situazione 1.6. Siano $x = (x^1, \dots, x^n)^t$, $y = (y^1, \dots, y^n)^t$ due punti in $\mathbb{R}^n$. Quando necessario (e lo sarà quasi sempre) supponiamo $n \geq 2$.

A partire dalla situazione 3.3 chiederemo inoltre che x ed y non siano diagonali, cioè che non abbiano coefficienti tutti uguali.

Definizione 1.7. La *media* $\overline{x}$ di x è definita come $\overline{x} := \dfrac{1}{n} \sum_{k=1}^{n} x^k$.

Osservazione 1.8. *La media è un operatore lineare; per $\lambda, \mu \in \mathbb{R}$ abbiamo quindi $\overline{\lambda x + \mu y} = \lambda \overline{x} + \mu \overline{y}$.*

Osservazione 1.9. *Per $\lambda \in \mathbb{R}$ si ha $\overline{\lambda^\diamond} = \lambda$. In particolare $\overline{\overline{x}^\diamond} = \overline{x}$.*

Dimostrazione. Infatti $\dfrac{\lambda + \ldots + \lambda}{n} = \lambda$.

Osservazione 1.10. $\|x, 1^\diamond\| = n\overline{x}$.

Dimostrazione. $\|x, 1^\diamond\| = x^1 + \ldots + x^n$.

Corollario 1.11. $x \perp 1^\diamond \iff \overline{x} = 0$.

I vettori che hanno media 0 sono quindi esattamente quei vettori che sono ortogonali alla retta diagonale; essi formano l'iperpiano $\mathbb{R}^{\diamond\perp}$ normale alla retta diagonale.

Matrici ausiliarie

Definizione 2.1. Denotiamo con δ la matrice identica in $\mathbb{R}^n_n$.

Definizione 2.2. $1^\square \in \mathbb{R}^n_n$ sia la matrice quadratica $n \times n$ i cui coefficienti sono tutti uguali ad 1:

$$1^\square := \begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$

Per $\lambda \in \mathbb{R}$ sia $\lambda^\square := \lambda 1^\square$.

Definiamo $M := (1/n)1^\square = (1/n)^\square$. Anche M è naturalmente una matrice quadratica $n \times n$.

Esempio 2.3. Per $n = 2$ quindi $M = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$, e per $n = 3$ $M = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$.

Osservazione 2.4. *$Mx = \overline{x}^\diamond$ e quindi $\overline{Mx} = \overline{x}$.*

Dimostrazione. Infatti

$$Mx = \frac{1}{n}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix} = \frac{1}{n}\begin{pmatrix} x^1 + \dots + x^n \\ \vdots \\ x^1 + \dots + x^n \end{pmatrix} = \begin{pmatrix} \overline{x} \\ \vdots \\ \overline{x} \end{pmatrix} = \overline{x}^\diamond$$

Corollario 2.5. $M1^\diamond = 1^\diamond$.

Dimostrazione. Per l'osservazione 2.4 e usando l'osservazione 1.9 abbiamo $M1^\diamond = \overline{1^\diamond}^\diamond = 1^\diamond$.

Osservazione 2.6. $(1^\square)^2 = n^\square = \begin{pmatrix} n & \dots & n \\ \vdots & & \vdots \\ n & \dots & n \end{pmatrix}$.

Dimostrazione. Chiaro da

$$\begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix} = \begin{pmatrix} n & \dots & n \\ \vdots & & \vdots \\ n & \dots & n \end{pmatrix}$$

Corollario 2.7. $M^2 = M$.

Dimostrazione. Per l'osservazione 2.6 abbiamo

$$M^2 = \frac{1}{n^2}(1^\square)^2 = \frac{1}{n^2}n^\square = (1/n)^\square = M$$

La centralizzazione

Definizione 2.8. La matrice $C := \delta - M$ si chiama la *matrice centralizzante* (di dimensione n).

Esempio 2.9. Per $n = 2$ quindi $M = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$, e per $n = 3$ $M = \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}$.

Corollario 2.10. $C^2 = C$.

Dimostrazione. Infatti dal corollario 2.7 segue

$$(\delta - M)^2 = \delta - 2M + M^2 = \delta - 2M + M = \delta - M$$

Osservazione 2.11. $MC = CM = 0$.

Dimostrazione. Utilizzando il corollario 2.7 abbiamo

$$MC = M(\delta - M) = M - M^2 = 0$$
$$CM = (\delta - M)M = M - M^2 = 0$$

Definizione 2.12. Il vettore $Cx = x - Mx$ si chiama la *centralizzazione* di x. Per definizione quindi $x = Cx + Mx$.

Proposizione 2.13. $\overline{Cx} = 0$.

Dimostrazione. $\overline{Cx} = \overline{x - Mx} = \overline{x} - \overline{Mx} = \overline{x} - \overline{x} = 0$.

Abbiamo usato la linearità della media (osservazione 1.8) e l'osservazione 2.4.

Potremmo anche utilizzare l'osservazione 2.11 al posto dell'osservazione 1.8 ; infatti $\overline{Cx}^\diamond = MCx = 0$ implica $\overline{Cx} = 0$.

Corollario 2.14. $CCx = Cx$.

Dimostrazione. Ciò segue direttamente dal corollario 2.10.

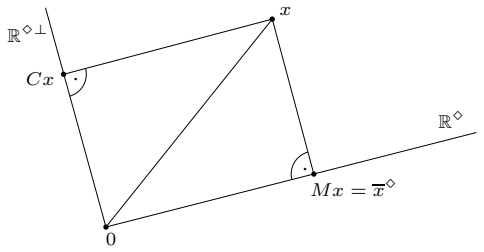
Corollario 2.15. *$Cx \perp 1^\diamond$ e quindi anche $Cx \perp Mx$.*

Dimostrazione. Proposizione 2.13 e corollario 1.11.

Teorema 2.16. *Cx è la proiezione ortogonale di x sull'iperpiano $\mathbb{R}^{\diamond\perp}$, mentre Mx è la proiezione ortogonale di x sulla retta diagonale.*

Dimostrazione. $Cx \in \mathbb{R}^{\diamond\perp}$ per il corollario 2.15, mentre è chiaro che $Mx = \overline{x}^\diamond$ appartiene alla retta diagonale.

Sia $\|v, 1^\diamond\| = 0$. Allora $\|v, x - Cx\| = \|v, \overline{x}^\diamond\| = 0$.
Infine $\|x - Mx, 1^\diamond\| = \|Cx, 1^\diamond\| = 0$.



Deviazione standard e varianza

Definizione 2.17. La *deviazione standard* s_x di x è definita da

$$s_x := \frac{|Cx|}{\sqrt{n-1}}$$

s_x^2 si chiama la *varianza* di x; abbiamo quindi $s_x^2 = \dfrac{|Cx|^2}{n-1}$.

La *covarianza* s_{xy} di x ed y è definita da $s_{xy} := \dfrac{\|Cx, Cy\|}{n-1}$.

Abbiamo in particolare $s_{xx} = s_x^2$.

Lemma 2.18. *Valgono le uguaglianze*

$$\|Cx, Cy\| = \|Cx, y\| = \|x, y\| - n\overline{x}\,\overline{y}$$

Da esse seguono le relazioni

$$s_{xy} = \frac{\|x, y\| - n\overline{x}\,\overline{y}}{n-1}$$
$$|Cx|^2 = |x|^2 - n\overline{x}^2$$
$$s_x^2 = \frac{|x|^2 - n\overline{x}^2}{n-1}$$

Queste formule sono usate molto spesso.

Dimostrazione. Per il corollario 2.15 e l'osservazione 1.10 abbiamo

$$\|Cx, Cy\| = \|Cx, y - \overline{y}^\diamond\| = \|Cx, y\| = \|x - \overline{x}^\diamond, y\|$$
$$= \|x, y\| - \overline{x}\|1^\diamond, y\| = \|x, y\| - n\overline{x}\,\overline{y}$$

Corollario 2.19. *Possiamo così calcolare le lunghezze delle proiezioni di x sull'iperpiano $\mathbb{R}^{\diamond\perp}$ e sulla retta diagonale $\mathbb{R}^\diamond$:*

$$|Cx| = |x - Mx| = \sqrt{|x|^2 - n\overline{x}^2}$$
$$|Mx| = |\overline{x}|\sqrt{n}$$

Le normalizzazioni $\widehat{x}$ ed $\breve{x}$

Osservazione 3.1. Sia $v \in \mathbb{R}^n$ e $v \neq 0$. Allora il vettore $\dfrac{v}{|v|}$ possiede lunghezza 1 e mostra naturalmente nella stessa direzione di v.

Osservazione 3.2. *Sono equivalenti:*

(1) x *è diagonale.*
(2) $x \in \mathbb{R}^{\diamond}$.
(3) $x = \overline{x}^{\diamond}$.
(4) $x = Mx$.
(5) $Cx = 0$.
(6) $s_x = 0$.

Situazione 3.3. Assumiamo da ora in avanti che x ed y non siano diagonali e quindi $Cx \neq 0$, $Cy \neq 0$. È chiaro che ciò implica che $n \geq 2$.

Dall'osservazione 3.2 vediamo anche che $s_x > 0$ ed $s_y > 0$.

Definizione 3.4. Il vettore $\widehat{x} := \dfrac{Cx}{|Cx|}$ si chiama la *normalizzazione geometrica* di x. In statistica si considera anche il vettore $\breve{x} := \dfrac{Cx}{s_x}$, che chiameremo la *normalizzazione statistica* di x.

Nota 3.5. $\breve{x} = \sqrt{n-1}\,\widehat{x}$.

$\breve{x}$ si distingue quindi da $\widehat{x}$ solo per il fattore $\sqrt{n-1}$. Le considerazioni geometriche che seguono potrebbero perciò essere eseguite anche con $\breve{x}$, risultano però più trasparenti e le formule che si ottengono più semplici, se si usa $\widehat{x}$.

Dimostrazione. Abbiamo $\breve{x} = \dfrac{Cx}{s_x} = \dfrac{Cx}{|Cx|}\dfrac{|Cx|}{s_x} = \widehat{x}\dfrac{|Cx|}{s_x}$.

Ma per la definizione 2.17 vale $\dfrac{|Cx|}{s_x} = \sqrt{n-1}$.

Osservazione 3.6. *$\widehat{x}$ ed $\breve{x}$ sono vettori paralleli a Cx, perciò* $\overline{\widehat{x}} = \overline{\breve{x}} = 0$.

Osservazione 3.7. *Sia $v \in \mathbb{R}^n$, $v \neq 0$ e $\overline{v} = 0$. Allora* $\widehat{x} = \dfrac{v}{|v|}$.

Corollario 3.8. $\widehat{\widehat{x}} = \widehat{x}$.

Dimostrazione. Ciò segue dalle osservazioni 3.6 e 3.7 perché $|\widehat{x}| = 1$.

Osservazione 3.9. *Sia $\alpha > 0$. Allora* $\widehat{\alpha x} = \widehat{x}$.

Dimostrazione. Infatti $C\alpha x = \alpha Cx$, per cui

$\widehat{\alpha x} = \dfrac{C\alpha x}{|C\alpha x|} = \dfrac{\alpha Cx}{|\alpha||Cx|} = \widehat{x}$, perché $|\alpha| = \alpha$.

Corollario 3.10. $\widehat{Cx} = \widehat{\breve{x}} = \widehat{x}$.

Dimostrazione. Dalla definizione 3.4 vediamo che Cx e $\breve{x}$ si distinguono da $\widehat{x}$ solo per i fattori positivi $|Cx|$ risp. s_x.

Esempio 3.11. Sia $x = \begin{pmatrix} 13 \\ 1 \\ 5 \\ 2 \\ 9 \end{pmatrix}$. Allora

$$\overline{x} = \frac{13+1+5+2+9}{5} = \frac{30}{5} = 6$$

quindi

$$Cx = \begin{pmatrix} 13 \\ 1 \\ 5 \\ 2 \\ 9 \end{pmatrix} - \begin{pmatrix} 6 \\ 6 \\ 6 \\ 6 \\ 6 \end{pmatrix} = \begin{pmatrix} 7 \\ -5 \\ -1 \\ -4 \\ 3 \end{pmatrix}$$

e $|Cx| = \sqrt{49+25+1+16+9} = \sqrt{100} = 10$, per cui

$$\widehat{x} = \frac{1}{10}Cx = \begin{pmatrix} 0.7 \\ -0.5 \\ -0.1 \\ -0.4 \\ 0.3 \end{pmatrix} \text{ e } \breve{x} = \sqrt{4}\,\widehat{x} = 2\widehat{x} = \begin{pmatrix} 1.4 \\ -1 \\ -0.2 \\ -0.8 \\ 0.6 \end{pmatrix}$$

Osservazione 3.12. Dal corollario 2.19 otteniamo la decomposizione ortonormale

$$x = \sqrt{|x|^2 - n\overline{x}^2}\cdot\widehat{x} + \overline{x}\sqrt{n}\cdot(1/\sqrt{n})^{\diamond}$$

I vettori $\widehat{x}$ e $(1/\sqrt{n})^{\diamond}$ sono ortogonali tra di loro e possiedono lunghezza 1. Si osservi che il secondo non dipende da x.

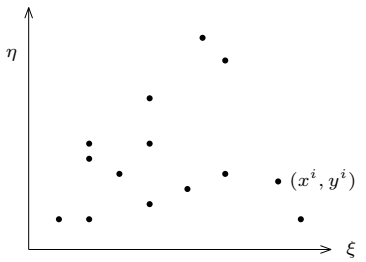
La formula mostra molto bene come $\overline{x}\sqrt{n}$ sia la distanza segnata di x dall'iperpiano $\mathbb{R}^{\diamond\perp}$ dei vettori di media 0.

x e $(1/\sqrt{n})^{\diamond}$ sono separati da questo iperpiano se e solo se $\overline{x} < 0$.

La retta di regressione

I *modelli lineari* sono impiegati con successo in molte indagini statistiche; questo capitolo è dedicato al caso più semplice, la rappresentazione di una dipendenza approssimativamente lineare di y da x mediante una retta di regressione.

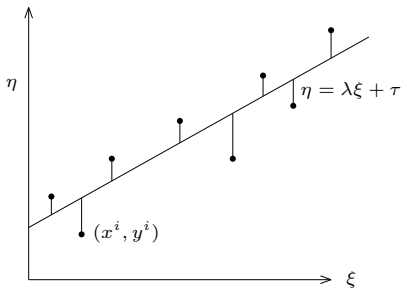
Nota 3.13. In statistica spesso in un primo momento sono dati n punti $(x^1, y^1), \ldots, (x^n, y^n)$ nel piano $\mathbb{R}_2$, da cui, secondo il principio di dualità, possiamo formare i vettori $x, y \in \mathbb{R}^n$. Avendo così già assegnato le lettere x e y, denotiamo le coordinate nel piano con ξ ed η.



Cerchiamo adesso di rappresentare (più precisamente di approssimare) i valori y^i mediante una funzione lineare degli x_i, cioè di determinare numeri reali λ e τ tali da minimizzare gli errori $y^i - (\lambda x^i + \tau)$ nel senso che l'espressione

$$F(\lambda, \tau) := \sum_{i=1}^{n} (y^i - (\lambda x^i + \tau))^2$$

sia minima (principio dei *minimi quadrati* di Gauss).



A questo scopo si possono porre uguali a zero le derivate parziali $\dfrac{\partial F}{\partial \lambda}$ e $\dfrac{\partial F}{\partial \tau}$, ottenendo così un sistema lineare in λ e τ che, nella nostra ipotesi che x non sia diagonale, possiede un'unica soluzione (λ, τ).

La retta determinata dall'equazione $\eta = \lambda\xi + \tau$ si chiama la *retta di regressione* degli y^i rispetto agli x^i (o di y rispetto ad x).

Nel seguito useremo (λ, τ) sia per denotare questa soluzione che per parametri generici variabili; sarà chiaro dal contesto quale dei due significati è usato.

Vogliamo adesso invece dedurre la retta di regressione senza fare uso del calcolo differenziale in modo puramente geometrico. Lavoriamo in $\mathbb{R}^n$ con x, y definiti come finora, nonostante che la retta di regressione sia una retta in $\mathbb{R}_2$ riferita ai punti (x^i, y^i).

Osservazione 3.14. Nella situazione della nota 3.134 abbiamo

$$F(\lambda, \tau) = |y - (\lambda x + \tau^{\diamond})|^2$$

Dobbiamo quindi scegliere λ e τ in modo da minimizzare la lunghezza del vettore $y - (\lambda x + \tau^{\diamond})$.

I coefficienti della retta di regressione

Proposizione 4.1. *E sia un sottospazio vettoriale di $\mathbb{R}^n$ ed $e_1, \dots, e_s$ una base ortogonale di E. Siano $y \in \mathbb{R}^n$ e p la proiezione ortogonale di y su E. Allora*

$$p = \alpha_1 e_1 + \dots + \alpha_s e_s$$

con gli α_k (naturalmente univocamente determinati) dati da

$$\alpha_k = \frac{\|y, e_k\|}{|e_k|^2}$$

Questa formula mostra in particolare che ogni sommando $p_k = \alpha_k e_k$ è la proiezione ortogonale di y sulla retta $\mathbb{R}e_k$ generata da e_k.
p si ottiene come $p = p_1 + \dots + p_s$.

Dimostrazione. $y - p$ deve essere ortogonale ad e_k per ogni k e quindi deve valere $\|y - p, e_k\| = 0$ o, equivalentemente,

$$\|y, e_k\| = \|p, e_k\|$$

per $k = 1, \dots, m$. Per l'ortogonalità degli e_j abbiamo però

$$\|p, e_k\| = \|\alpha_k e_k, e_k\| = \alpha \|e_k, e_k\|$$

cosicché $\alpha_k \|e_k, e_k\| = \|y, e_k\|$ e ciò implica l'enunciato.

Nota 4.2. Ci mettiamo di nuovo nella situazione della nota 3.13. Siccome per ipotesi x non si trova sulla retta $\mathbb{R}^\diamond$, i punti x ed $1^\diamond$ generano un piano $P_x \subset \mathbb{R}^n$:

$$P_x = \{\lambda x + \tau^\diamond \mid \lambda, \tau \in \mathbb{R}\}$$

in cui λ e τ per ogni punto di P_x sono univocamente determinati. In particolare sono univocamente determinati i parametri λ e τ corrispondenti alla proiezione ortogonale p di y su P_x. Ma p è proprio il punto per il quale $F(\lambda, \tau)$ è minimale.

D'altra parte anche $Cx = x - \overline{x}1^\diamond$ appartiene a P_x e dal corollario 2.15 segue adesso che Cx e $1^\diamond$ formano una base ortogonale di P_x, quindi, per la proposizione 4.1,

$$p = p_1 + p_2$$

dove p_1 è la proiezione ortogonale di y sulla retta generata da Cx e p_2 la proiezione ortogonale di y sulla retta generata da $1^\diamond$. Dal teorema 2.16 sappiamo però anche che $p_2 = My$. Abbiamo quindi, con un $\alpha \in \mathbb{R}$ che naturalmente è determinato dalla formula della proposizione 4.1,

$$\begin{aligned} p &= \alpha Cx + My \\ &= \alpha(x - \overline{x}^\diamond) + \overline{y}^\diamond \\ &= \alpha x + (\overline{y} - \alpha\overline{x})^\diamond \end{aligned}$$

Ciò mostra che

$$\begin{aligned} \lambda &= \alpha \\ \tau &= \overline{y} - \lambda\overline{x} \end{aligned}$$

Notiamo che a questo punto abbiamo $p = \overline{y}^\diamond + \lambda Cx$.

Dobbiamo ancora calcolare λ. Per la proposizione 4.1 e usando il lemma 2.18 abbiamo

$$\lambda = \frac{\|y, Cx\|}{|Cx|^2} = \frac{\|Cx, Cy\|}{|Cx|^2} = \frac{\|Cx, Cy\|}{|Cx||Cy|} \cdot \frac{|Cy|}{|Cx|}$$

Se poniamo

$$r_{xy} := \frac{\|Cx, Cy\|}{|Cx||Cy|} = \|\widehat{x}, \widehat{y}\|$$

abbiamo infine

$$\boxed{\begin{aligned} \lambda &= r_{xy} \frac{|Cy|}{|Cx|} \\ \tau &= \overline{y} - \lambda\overline{x} \end{aligned}}$$

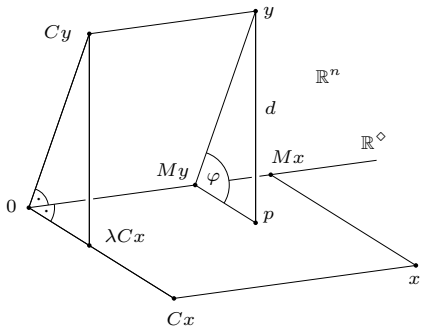
La retta di regressione di y rispetto ad x possiede quindi l'equazione

$$\eta = \lambda\xi + \tau$$

con λ e τ come sopra.

Definizione 4.3. Il rapporto r_{xy} definito nella nota 4.2 si chiama il *coefficiente di correlazione* tra x ed y e verrà studiato in dettaglio nel prossimo capitolo.

Dai corsi di Geometria sappiamo che r_{xy} non è altro che il coseno dell'angolo φ tra Cx e Cy.



Si osservi che, nonostante si tratti di un disegno in $\mathbb{R}^n$, questa figura è realistica nel senso che la configurazione è tutta contenuta nello spazio (al massimo) 3-dimensionale generato dai tre vettori $1^\diamond$, Cx e Cy. È indicato il vettore dei residui d (pagina 5).

Osservazione 4.4. Siccome $\tau = \overline{y} - \lambda\overline{x}$, l'equazione $\eta = \lambda\xi + \tau$ per la retta di regressione diventa $\eta = \lambda\xi + \overline{y} - \lambda\overline{x}$ e può perciò essere scritta nella forma

$$\eta - \overline{y} = \lambda(\xi - \overline{x})$$

Essa passa quindi per il baricentro $(\overline{x}, \overline{y})$ dei punti (x^i, y^i). Inoltre

$$\lambda = r_{xy} \frac{|Cy|}{|Cx|} = r_{xy} \frac{|y - My|}{|x - Mx|}$$

cosicché l'equazione assume la forma

$$\frac{\eta - \overline{y}}{|y - My|} = r_{xy} \frac{\xi - \overline{x}}{|x - Mx|}$$

Nota 4.5. Siccome $\widehat{x}$ e $\widehat{y}$ si distinguono da Cx e Cy solo per fattori positivi, è chiaro che $\widehat{x}$ e $\widehat{y}$ racchiudono lo stesso angolo come Cx e Cy; lo stesso vale per $\check{x}$ e $\check{y}$.

In particolare vediamo che il coefficiente di correlazione può anche essere definito come il coseno dell'angolo tra $\widehat{x}$ e $\widehat{y}$ e che quindi per il corollario 3.10 il coefficiente di correlazione non cambia se sostituiamo x ed y con le loro normalizzazioni geometriche o statistiche o con le loro centralizzazioni.

Osservazione 4.6. Dalla definizione 2.17 vediamo che $\lambda = r_{xy} \dfrac{s_y}{s_x}$.

Osservazione 4.7. Siccome $C^2 = C$ (corollari 2.10 e 2.14), dalle formule nella nota 4.2 si vede che se sostituiamo x ed y con Cx e Cy, il coefficiente λ nella retta di regressione non cambia, mentre il coefficiente τ diventa uguale a 0, perché Cx e Cy hanno media zero.

Nota 4.8. Nei calcoli a mano o per ragioni numeriche conviene talvolta effettuare una trasformazione affine dei dati. Come si comportano i coefficienti della retta di regressione?

Nella prima tabella che segue si vede facilmente che x ed y possono essere trasformate nei dati $\widetilde{x}$ e $\widetilde{y}$ elencati nella seconda tabella:

x	y
11250	67
11280	66
11300	61
11200	68
11360	64

$\widetilde{x}$	$\widetilde{y}$
5	6
8	5
10	0
0	7
16	3

Abbiamo quindi effettuato la trasformazione $\widetilde{x} = x/10 - 1120^\diamond$, $\widetilde{y} = y - 61^\diamond$. Nel caso generale di una trasformazione

$$\widetilde{x} = ax + b^\diamond, \widetilde{y} = cy + d^\diamond$$

si trova facilmente $\widetilde{\lambda} = (c/a)\lambda$ e $\widetilde{\tau} = c\tau + d - (bc/a)\lambda$, da cui $\lambda = (a/c)\widetilde{\lambda}$ e $\tau = (\widetilde{\tau} - d + b\widetilde{\lambda})/c$. Nel nostro esempio abbiamo $\lambda = \widetilde{\lambda}/10$ e $\tau = \widetilde{\tau} + 61 - 1120\widetilde{\lambda}$.

Osservazioni generali

Applichiamo la teoria a una tabella che si trova a pagina 263 dell'ottimo libro di Kreyszig. La tabella contiene nella colonna degli x^i le densità molteplicate per 10 di esemplari di minerali di ematite; gli y^i sono i contenuti percentuali di ferro.

x	y
28	27
29	23
30	30
31	28
32	30
32	32
32	34
33	33
34	30

Facendo i conti, troviamo $\lambda = 1.21, \tau = -8.01$.

---

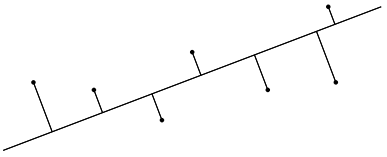
x ed y siano dati dalla tabella

x	y
1	0
0	1
-1	0
0	-1

In questo esempio $\overline{x} = \overline{y} = 0$, quindi $Cx = x$, $Cy = y$ e $\tau = 0$. Inoltre $Cx \perp Cy$, per cui $r_{xy} = 0$ e quindi anche $\lambda = 0$. La retta di regressione è perciò l'ascisse reale $\eta = 0$.

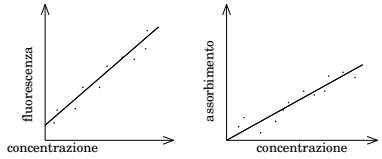
Nota 5.1. Siccome $\lambda = r_{xy}\dfrac{|Cy|}{|Cx|}$ e siccome per ipotesi $Cy \neq 0$,
è chiaro che la retta di regressione è parallela all'ascissa reale, come nell'ultimo esempio, se e solo se il coefficiente di correlazione si annulla, e ciò accade se e solo se $Cx \perp Cy$.

L'uso della retta di regressione è giustificato soprattutto quando i valori x^i e y^i rappresentano misurazioni di variabili tra le quali è nota l'esistenza di un legame *lineare* che però è stato confuso da errori nella misurazione degli y^i. In questo caso si può assumere che la retta di regressione rappresenti questo legame lineare. Se coesistono errori di misurazione in entrambe le variabili, è preferibile la *regressione ortogonale* mediante proiezioni ortogonali su una retta (invece di proiezioni parallele all'asse y); essa appartiene all'*analisi delle componenti principali* che verrà trattata più avanti.



Regressione ortogonale

In chimica analitica si incontrano spesso leggi lineari che possono essere caratterizzate mediante regressione e correlazione (Otto, Doerffel). Si cerca ad esempio di calcolare la dipendenza spesso lineare dei segnali di misurazione dai parametri chimici (*curve di calibrazione*).

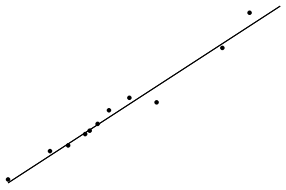


In una *serie temporale* la variabile x è interpretata come il tempo, y come una variabile dipendente dal tempo. Non raramente si osserva una tendenza (in inglese *trend*) lineare a cui si sovrappongono oscillazioni più o meno periodiche e che può essere rapresentata mediante una retta di regressione.

I parametri λ e τ dell'analisi regressionale, calcolati algebricamente, dovrebbero essere stimati, soprattutto se vengono utilizzati a scopi interpolatori. Per fare ciò bisogna o fare ipotesi sulla distribuzione statistica delle variabili casuali corrispondenti alle variabili empiriche x ed y (ad esempio assumendo una distribuzione normale) oppure usare metodi nonparametrici. Non sempre è sicuro che veramente esiste un legame di base (ad es. fisico-chimico) lineare; in questi casi anche la linearità della dipendenza deve essere verificata con metodi statistici.

Legami lineari si osservano spesso nei livelli d'acqua in due postazioni idrometriche distanti allo stesso fiume. Un esempio dal trattato di idrologia di Maniak, pag. 200, leggermente modificato:

x	y
309	193
302	187
283	174
443	291
298	184
319	205
419	260
361	212
267	169
337	216
230	144



I livelli nelle due postazioni sono indicati in cm. Si calcola $\lambda = 0.65$, $\tau = -8.6$.

Il modello con una variabile indipendente ξ nelle applicazioni pratiche è spesso troppo semplice; modelli molti più efficaci si ottengono con *regressioni lineari multiple* della forma

$$\eta = \lambda_1\xi_1 + \ldots + \lambda_k\xi_k + \tau$$

Tali modelli sono già molto generali e vengono usati in molti problemi ingegneristici o econometrici o ad esempio nell'idrologia nella prognosi dei livelli d'acqua, in modo simile alla regressione semplice che abbiamo visto nell'ultimo esempio.

K. Doerffel: Statistik in der analytischen Chemie. Grundstoffindustrie 1990.

E. Kreyszig: Statistische Methoden und ihre Anwendungen. Vandenhoeck 1975.

U. Maniak: Hydrologie und Wasserwirtschaft. Springer 1997.

M. Otto: Chemometrics. VCH 1999.

Analisi dei residui

Nell'*analisi dei residui* di una retta di regressione si studiano le differenze (i *residui*)

$$d^i = y^i - (\lambda x^i + \tau)$$

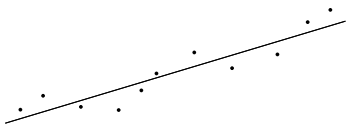
Si ottengono tra l'altro indicazioni per un eventuale possibile miglioramento del modello di regressione.

Lavoriamo di nuovo in $\mathbb{R}^n$ e introduciamo il vettore dei residui

$$d := y - (\lambda x + \tau^{\diamond}) = y - p$$

d è quindi semplicemente il vettore che congiunge la proiezione ortogonale p di y su P_x con y; cfr. la figura nella definizione 4.3.

Analizzando il vettore dei residui si trova spesso che esso può essere decomposto in più componenti; in questo caso si dovrebbe tentare una regressione multipla. Una rappresentazione grafica dei residui permette talvolta di riconoscere fenomeni di periodicità che possono suggerire l'utilizzo di un nuovo modello non lineare.



L'analisi dei residui è particolarmente utile nella ricerca di *errori sistematici* (Doerffel, 171-177, Otto, 207-215).

II. IL COEFFICIENTE DI CORRELAZIONE

Il prodotto scalare

Situazione 6.1. Siano $x, y \in \mathbb{R}^n$ con $x = (x^1, \dots, x^n)^t$, $y = (y_1, \dots, y_n)^t$. Supponiamo di nuovo che x ed y non siano diagonali. Possiamo allora formare il coefficiente di correlazione

$$r := r_{xy} = \frac{\|Cx, Cy\|}{|Cx||Cy|} = \|\widehat{x}, \widehat{y}\|$$

già introdotto nella definizione 4.3.

$\varphi, \lambda, \tau, p$ sono definiti come a pagina 3.

Nota 6.2. L'equazione $\|x, 1^\diamond\| = n\overline{x}$ dell'osservazione 1.10, benché immediata nella dimostrazione, stabilisce un importante legame tra un concetto statistico, la media $\overline{x}$, e un concetto geometrico, il prodotto scalare.

Il coefficiente di correlazione è definito mediante un prodotto scalare. Il prodotto scalare di due vettori $u, v \in \mathbb{R}^n$ è a sua volta profondamente legato alla lunghezza $|u + v|$ della somma di due vettori oppure anche alla lunghezza $|u - v|$ della differenza. Abbiamo infatti

$$\begin{aligned} |u+v|^2 &= \sum_{k=1}^n (u^k + v^k)^2 = \sum_{k=1}^n (u^k)^2 + \sum_{k=1}^n (v^k)^2 + 2\sum_{k=1}^n u^k v^k \\ &= |u|^2 + |v|^2 + 2\|u, v\| \end{aligned}$$

e similmente $|u - v|^2 = |u|^2 + |v|^2 - 2\|u, v\|$.

I due punti u e v formano insieme all'origine 0 un triangolo (eventualmente degenerato) i cui lati hanno le lunghezze $|u|$, $|v|$ e $|u - v|$.

Assumiamo che il triangolo non sia degenerato e sia α l'angolo opposto al lato di lunghezza $|u - v|$. Per il teorema del coseno abbiamo

$$|u - v|^2 = |u|^2 + |v|^2 - 2|u||v| \cos\alpha$$

da cui $\|u, v\| = |u||v| \cos\alpha$, come abbiamo già osservato a pagina 4.

Il coefficiente di correlazione di x ed y, nonostante il nome prometta molto di più, è essenzialmente un parametro che lega $\widehat{x}$ ed $\widehat{y}$ ad $\widehat{x} + \widehat{y}$ ed $\widehat{x} - \widehat{y}$.

Corollario 6.3. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$|\alpha u + \beta v|^2 = \alpha^2|u|^2 + \beta^2|v|^2 + 2\alpha\beta\|u, v\|$$

Dimostrazione. Per la nota 6.2 e la bilinearità del prodotto scalare abbiamo

$$\begin{aligned} |\alpha u + \beta v|^2 &= |\alpha u|^2 + |\beta v|^2 + 2\|\alpha u, \beta v\| \\ &= \alpha^2|u|^2 + \beta^2|v|^2 + 2\alpha\beta\|u, v\| \end{aligned}$$

Lemma 6.4. *Siano $u, v \in \mathbb{R}^n$ vettori di lunghezza 1, cioè $|u| = |v| = 1$. Allora*

$$\|u, v\| = 1 - \frac{1}{2}|u - v|^2$$

Dimostrazione. Per la nota 6.2 abbiamo

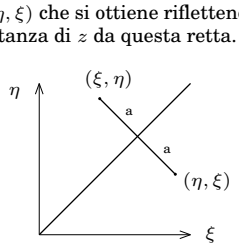
$$|u - v|^2 = |u|^2 + |v|^2 - 2\|u, v\| = 2 - 2\|u, v\|$$

per cui $2\|u, v\| = 2 - |u - v|^2$. Ciò implica l'enunciato.

Nota 6.5. $\dfrac{(\xi - \eta)^2}{2}$ è il quadrato della distanza del punto (ξ, η) dalla retta $\eta = \xi$.

Dimostrazione. Consideriamo un punto $z = (\xi, \eta)$ nel piano e il punto $z' = (\eta, \xi)$ che si ottiene riflettendo z alla retta $\eta = \xi$.

a sia la distanza di z da questa retta.



Allora $z - z' = (\xi - \eta, \eta - \xi)$, per cui $(2a)^2 = |z - z'|^2 = 2(\xi - \eta)^2$, cosicché $a^2 = \frac{1}{2}(\xi - \eta)^2$.

Nota 6.6. Siano $u, v \in \mathbb{R}^n$ vettori di lunghezza 1. Per $i = 1, \dots, n$ sia a^i la distanza del punto (u^i, v^i) dalla retta $\eta = \xi$ in $\mathbb{R}_2$. Allora

$$\|u, v\| = 1 - \sum_{i=1}^n (a^i)^2$$

Dimostrazione. Dalla nota 6.5 sappiamo che $(a^i)^2 = \frac{1}{2}(u^i - v^i)^2$. Per il lemma 6.4

$$\|u, v\| = 1 - \frac{1}{2}|u - v|^2 = 1 - \sum_{i=1}^n \frac{1}{2}(u^i - v^i)^2 = 1 - \sum_{i=1}^n (a^i)^2$$

Algebra della varianza

Proposizione 6.7. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$s^2_{\alpha u + \beta v} = \alpha^2 s_u^2 + \beta^2 s_v^2 + 2\alpha\beta s_{uv}$$

Dimostrazione. Usando il corollario 6.3 abbiamo

$$\begin{aligned} s^2_{\alpha u + \beta v} &= \frac{|C(\alpha u + \beta v)|^2}{n-1} = \frac{|\alpha Cu + \beta Cv|^2}{n-1} \\ &= \alpha^2 \frac{|Cu|^2}{n-1} + \beta^2 \frac{|Cv|^2}{n-1} + 2\alpha\beta \frac{\|Cu, Cv\|}{n-1} \\ &= \alpha^2 s_u^2 + \beta^2 s_v^2 + 2\alpha\beta s_{uv} \end{aligned}$$

Corollario 6.8. *Siano $u, v \in \mathbb{R}^n$. Allora*

$$s^2_{u+v} = s_u^2 + s_v^2 + 2s_{uv}$$

Osservazione 6.9. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$s_{\alpha u, \beta v} = \alpha\beta s_{uv}$$

Dimostrazione. Infatti

$$\begin{aligned} s_{\alpha u, \beta v} &= \frac{\|C(\alpha u), C(\beta v)\|}{n-1} = \frac{\|\alpha Cu, \beta Cv\|}{n-1} \\ &= \alpha\beta \frac{\|Cu, Cv\|}{n-1} = \alpha\beta s_{uv} \end{aligned}$$

Osservazione 6.10. *Siano $u, v \in \mathbb{R}^n$ ed $\alpha, \beta \in \mathbb{R}$. Allora*

$$s_{u+\alpha^\diamond, v+\beta^\diamond} = s_{uv}$$

In particolare $s_{u+\alpha^\diamond} = s_u$.

Dimostrazione. Abbiamo $C(u + \alpha^\diamond) = Cu + C\alpha^\diamond = Cu$ perché $C\alpha^\diamond = 0$; per la stessa ragione $C(v + \beta^\diamond) = Cv$.

Ciò implica l'enunciato.

Nota 6.11. La deviazione standard s_x è, secondo la def. 2.17, uguale alla lunghezza del vettore Cx moltiplicata con il fattore $1/\sqrt{n-1}$ che non dipende da x, ma solo da n. Essa è quindi effettivamente una misura per la deviazione dei dati x^i dalla loro media $\overline{x}$.

Più difficile è l'interpretazione della covarianza s_{xy} di due vettori di dati x ed y. Essa può essere scritta nella forma

$$s_{xy} = \frac{\|Cx, Cy\|}{n-1} = \frac{1}{n-1}|Cx||Cy| \cos\varphi$$

dove con φ, come nella def. 4.3, abbiamo denotato l'angolo tra Cx e Cy. Essa è quindi piuttosto un'informazione sulla posizione geometrica reciproca di Cx e Cy il cui significato statistico o causale è alquanto dubbio. Per questa ragione bisogna essere molto prudenti nelle interpretazioni della covarianza o del coefficiente di correlazione come vedremo anche in alcuni esempi in questo capitolo.

Il fatto che la covarianza assume invece una giustificata importanza nel caso di una distribuzione normale induce talvolta ad attribuirle un significato anche nel caso generale. Ma covarianza e correlazione non sono strumenti adatti a scoprire il *tipo* di legame tra vettori di dati, ma soltanto l'*intensità* di questo legame una volta che si è stabilito in altro modo il tipo di legame e che sia un legame a cui parametri lineari possono essere applicati.

Il coefficiente di correlazione

Corollario 7.1. $r = 1 - \sum_{i=1}^{n} (a^i)^2$

dove $(a^i)^2$ è il quadrato della distanza di $(\widehat{x}^i, \widehat{y}^i)$ dalla retta $\eta = \xi$ nel piano $\mathbb{R}_2$.

Questa è una della più importanti interpretazioni del coefficiente di correlazione.

Dimostrazione. Ciò segue dalla nota 6.6, perché $r = \|\widehat{x}, \widehat{y}\|$.

Proposizione 7.2. $r = \dfrac{s_{xy}}{s_x s_y}$.

Dimostrazione. Abbiamo

$$s_{xy} = \frac{\|Cx, Cy\|}{n-1}$$

$$r = \frac{\|Cx, Cy\|}{|Cx||Cy|} = \frac{\|Cx, Cy\|}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

Corollario 7.3. $r = 0 \iff s_{xy} = 0$.

Corollario 7.4. $r = 0 \iff s^2_{x+y} = s^2_x + s^2_y$.

Dimostrazione. Ciò segue dai corollari 7.3 e 6.8.

Corollario 7.5. *Siano $\alpha, \beta \in \mathbb{R} \setminus 0$. Allora*

$$r_{\alpha x, \beta y} = (\operatorname{sgn} \alpha\beta) \cdot r_{xy}$$

Dimostrazione. Usando l'osservazione 6.9 e la proposizione 6.7 dalla proposizione 7.2 abbiamo

$$r_{\alpha x, \beta y} = \frac{s_{\alpha x, \beta y}}{s_{\alpha x} s_{\beta y}} = \frac{\alpha\beta}{|\alpha||\beta|} \frac{s_{xy}}{s_x s_y}$$

Corollario 7.6. *Siano $\alpha, \beta \in \mathbb{R}$. Allora $r_{x+\alpha^\diamond, y+\beta^\diamond} = r_{xy}$.*

Dimostrazione. Ciò segue dalla proposizione 7.2 e dall'osservazione 6.10, oppure in modo geometrico dalla figura nella definizione 4.3.

Nota 7.7. Abbiamo visto nell'osservazione 4.4 che la retta di regressione di y rispetto ad x può essere scritta nella forma

$$\eta - \overline{y} = \lambda(\xi - \overline{x}) \text{ con } \lambda = r\frac{|Cy|}{|Cx|}.$$

Sostituituiamo adesso x ed y con $\widehat{x}$ ed $\widehat{y}$. Il coefficiente di correlazione non cambia e le medie sono uguali a 0. Inoltre

$$|C\widehat{x}| = |\widehat{x}| = 1$$
$$|C\widehat{y}| = |\widehat{y}| = 1$$

per cui l'equazione della retta di regressione di $\widehat{y}$ rispetto ad $\widehat{y}$ è semplicemente

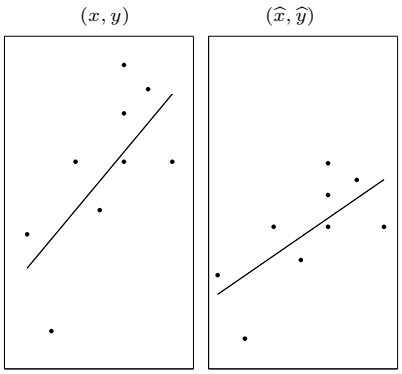
$$\eta = r\xi$$

Il coefficiente di correlazione è quindi la pendenza della retta di regressione di $\widehat{y}$ rispetto ad $\widehat{x}$.

Esempio 7.8. x ed y siano i dati relativi ai minerali di ematite della prima tabella a pagina 5. Calcolando le normalizzazioni geometriche otteniamo la tabella

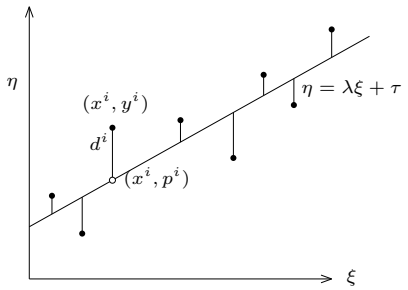
x	y	$\widehat{x}$	$\widehat{y}$
28	27	−0.59	−0.28
29	23	−0.41	−0.70
30	30	−0.22	0.04
31	28	−0.04	−0.18
32	30	0.14	0.04
32	32	0.14	0.25
32	34	0.14	0.46
33	33	0.33	0.35
34	30	0.51	0.04

Nelle figure sono indicate le rispettive rette di regressione. Sappiamo da pagina 5 che per x e y abbiamo $\lambda = 1.21$ e $\tau = -8.01$. Per $\widehat{x}$ ed $\widehat{y}$ dobbiamo calcolare il coefficiente di correlazione: troviamo $r = 0.69$.



Nota 7.9. Ricordiamo dalla nota 4.2 che $p = \lambda x + \tau^\diamond$.

Il vettore dei residui $d := y - p$ è stato introdotto a pagina 5. Le coordinate di $p = \lambda x + \tau^\diamond$ sono naturalmente $p^i = \lambda x^i + \tau$, i punti (x^i, p^i) sono quindi esattamente i punti sulla retta di regressione con ascisse uguale ad x^i. I residui sono $d^i = y^i - (\lambda x^i + \tau) = y^i - p^i$.



Proposizione 7.10. $|d|^2 = (1 - r^2)|Cy|^2$.

Dimostrazione. Nella figura della def. 4.3 vediamo che $\dfrac{d}{|Cy|} = |\sin\varphi| = \sqrt{1 - r^2}$ e ciò implica il risultato.

Osservazione 7.11. $-1 \le r \le 1$.

Dimostrazione. Sappiamo che $r = \cos\varphi$.

Corollario 7.12. *Sono equivalenti:*

(1) $d = 0$.
(2) I punti (x^i, y^i) si trovano tutti sulla retta di regressione di y rispetto ad x.
(3) $r^2 = 1$.
(4) $r = \pm 1$.

Dimostrazione. (1) $\iff$ (2): Chiaro.

(1) $\iff$ (3): Siccome $|Cy| \neq 0$, dalla proposizione 7.10 segue che $d = 0 \iff 1 - r^2 = 0$.

(3) $\iff$ (4): Chiaro.

Osservazione 7.13.
(1) $r = 1 \iff \widehat{x} = \widehat{y}$.
(2) $r = -1 \iff \widehat{x} = -\widehat{y}$.
(3) $r = 0 \iff \widehat{x} \perp \widehat{y} \iff Cx \perp Cy$.

Dimostrazione. $r = \cos\varphi$ e abbiamo già osservato che φ è anche l'angolo tra le normalizzazioni geometriche $\widehat{x}$ ed $\widehat{y}$.

Osservazione 7.14. *Sia $\overline{x} = 0$. Allora* $\|Cx, Cy\| = \|x, y\|$.

Dimostrazione. Per il lemma 2.18 abbiamo

$$\|Cx, Cy\| = \|Cx, y\| = \|x, y\|$$

perché $\overline{x} = 0$ implica $Cx = x$.

Corollario 7.15. *Sia $\overline{x} = 0$. Allora* $r = 0 \iff x \perp y$.

Decomposizione della varianza

Osservazione 8.1. *Sia* $u \in \mathbb{R}^n$. *Allora* $s_{Cu} = s_u$.

Dimostrazione. Ciò segue dall'osservazione 6.10.

Osservazione 8.2. $\overline{p} = \overline{y}$.

Dimostrazione. Dalla figura nel corollario 8.5 è chiaro che $My = \overline{y}^\diamond$ non è solo la proiezione ortogonale di y sulla retta $\mathbb{R}^\diamond$ (teorema 2.16), ma anche la proiezione ortogonale di p sulla stessa retta e ciò implica (ancora per il teorema 2.16) che $\overline{y}^\diamond = \overline{p}^\diamond$ e quindi $\overline{y} = \overline{p}$.

La dimostrazione analitica è altrettanto facile: Dalla nota 4.2 sappiamo che

$$p = \overline{y}^\diamond + \lambda Cx$$

Però $\overline{Cx} = 0$, per cui $\overline{p} = \overline{\overline{y}^\diamond} = \overline{y}$.

Corollario 8.3. $\overline{d} = 0$ *e quindi* $d = Cd$.

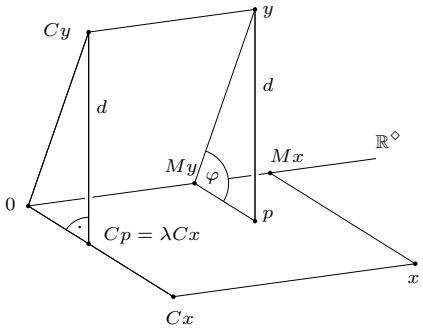
Dimostrazione. Ciò segue dall'osservazione 8.2 ed è evidente anche dalla figura nel corollario 8.5, da cui si vede che d è ortogonale a $1^\diamond$ e possiede quindi media 0 per il corollario 1.11.

Corollario 8.4. $Cp = \lambda Cx$.

Dimostrazione. Infatti $\lambda Cx = p - \overline{y}^\diamond = p - \overline{p}^\diamond = Cp$ per l'osservazione 8.2.

Corollario 8.5. $|Cy|^2 = |y - p|^2 + |Cp|^2 = |d|^2 + |Cp|^2$.

Usando il corollario 8.4 l'enunciato segue dalla figura - non è altro che il teorema di Pitagora applicato al triangolo a sinistra.



Proposizione 8.6. $|Cp|^2 = r^2|Cy|^2$.

Dimostrazione. Per il corollario 8.5 e la proposizione 7.10 abbiamo

$$|Cp|^2 = |Cy|^2 - |d|^2 = |Cy|^2 - (1 - r^2)|Cy|^2 = r^2|Cy|^2$$

Proposizione 8.7. $s_p^2 = \lambda^2 s_x^2$.

Dimostrazione. Dal corollario 8.4 abbiamo $Cp = \lambda Cx$. L'enunciato segue dall'osservazione 8.1 e dalla proposizione 6.7.

Teorema 8.8. $s_y^2 = s_p^2 + s_d^2 = \lambda^2 s_x^2 + s_d^2$.

Dimostrazione. Ciò segue dal corollario 8.5, perché dal corollario 8.3 sappiamo che $d = Cd$, per cui abbiamo

$$|Cy|^2 = (n-1)s_y^2 \quad |Cp|^2 = (n-1)s_p^2 \quad |d|^2 = |Cd|^2 = (n-1)s_d^2$$

Nota 8.9. Il teorema 8.8 è molto importante in statistica e costituisce una *decomposizione della varianza* di y nella somma tra la varianza di p, cioè la parte di s_y che deriva direttamente dalla regressione di y rispetto ad x, e la varianza di d, cioè la varianza del vettore dei residui.

s_d^2 perciò si chiama anche la *varianza residua* (di y rispetto ad x). La varianza di y è quindi uguale alla varianza dovuta alla regressione più la varianza residua.

Definizione 8.10. Il quoziente $\dfrac{s_p^2}{s_y^2} = \lambda^2 \dfrac{s_x^2}{s_y^2}$ dà una misura di quanto la regressione da sola determina la varianza di y e si chiama per questa ragione il *coefficiente di determinazione* (di y rispetto ad x).

Proposizione 8.11. *Il coefficiente di determinazione è uguale al quadrato del coefficiente di correlazione:* $\dfrac{s_p^2}{s_y^2} = r^2$.

Dimostrazione. Ciò segue direttamente dalla proposizione 8.7 e dall'equazione

$$\lambda = r\frac{s_y}{s_x}$$

che abbiamo visto nell'osservazione 4.6.

Nota 8.12. Per il corollario 7.12 il coefficiente di determinazione è uguale a 1 se e solo se i punti (x^i, y^i) si trovano tutti sulla retta di regressione di y rispetto ad x. Dalla proposizione 8.11 segue inoltre che il coefficiente di determinazione non cambia se scambiamo x ed y; infatti per definizione $r_{xy} = r_{yx}$.

Nota 8.13. Nelle ipotesi che abbiamo fatto nelle osservazioni che seguono la nota 5.1 le variabili x^i ed y^i hanno ruoli diversi. In situazioni in cui nessuna delle due variabili può essere considerata indipendente si può disegnare anche la retta di regressione

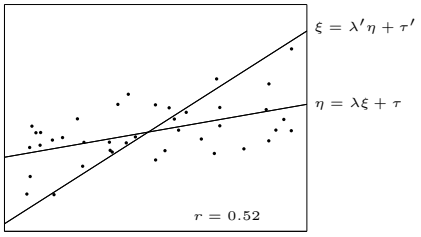
$$\xi = \lambda'\eta + \tau'$$

degli x^i rispetto agli y^i. Allora, siccome $r_{xy} = r_{yx} = r$, abbiamo

$$\lambda = r\frac{|Cy|}{|Cx|} \qquad \lambda' = r\frac{|Cx|}{|Cy|}$$

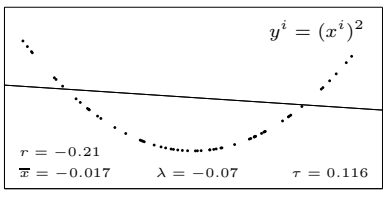
Da ciò segue $r^2 = \lambda\lambda'$.

$|r| = \sqrt{\lambda\lambda'}$ è quindi la *media geometrica* delle pendenze delle due rette di regressione.



Evitare interpretazioni causali

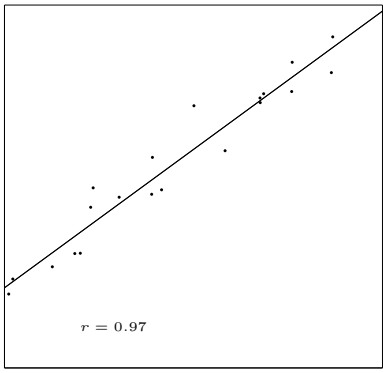
Nota 8.14. Abbiamo visto finora le più importanti interpretazioni del coefficiente di correlazione. Esse mostrano che si tratta di un concetto essenzialmente geometrico che dovrebbe essere quindi utilizzato solo in quei casi in cui i legami geometrici hanno un significato statistico per il problema che si studia. In particolare si dovrebbero evitare interpretazioni *causali*, anche in casi di correlazioni vicine a 1. Una correlazione uguale o vicina a 0 a sua volta non implica che non ci sono legami statistici o causali tra le variabili. Se ad esempio $\overline{x} = 0$ e con ogni punto (x^i, y^i) anche $(-x^i, y^i)$ appartiene ai dati (con la stessa molteplicità se presente più volte), per il corollario 7.15 il coefficiente di correlazione si annulla, anche quando sussiste un semplice legame funzionale tra le variabili, ad esempio ogni volta che $y^i = f(x^i)$, dove f è una funzione simmetrica, cioè tale che $f(\xi) = f(-\xi)$.



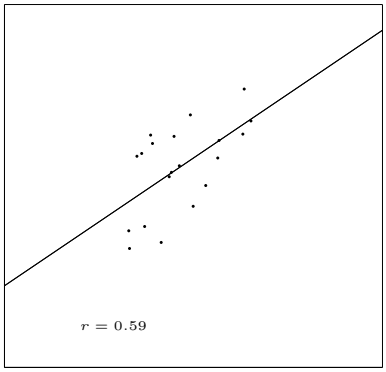
In questo caso la retta di regressione è data da $\eta = \overline{y}$, come segue dalla relazione $\tau = \overline{y} - \lambda\overline{x}$.

Un coefficiente di correlazione nullo non significa quindi una mancanza di legami causali tra x ed y, ma esprime piuttosto una forma di *simmetria*.

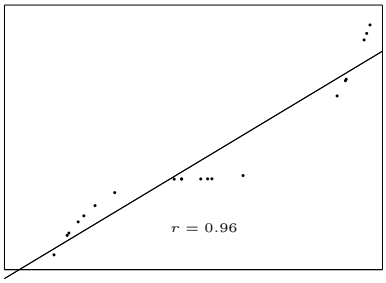
Esempi commentati



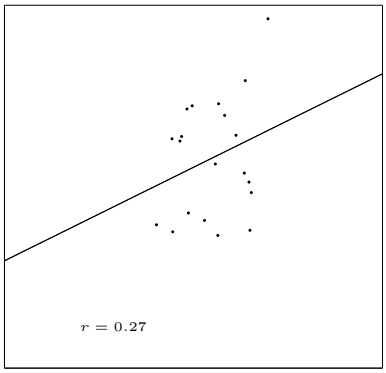
In questo caso y sembra veramente dipendere in modo lineare da x; la retta di regressione può essere utilizzata correttamente come legge che lega le due variabili.

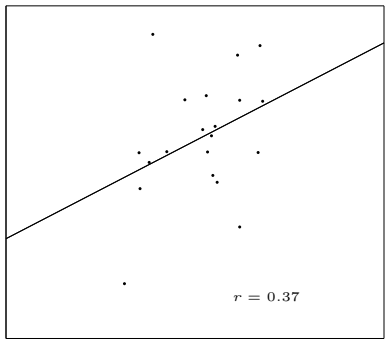


Questo caso è simile al precedente con il coefficiente di correlazione che esprime correttamente il più debole legame rispetto al caso precedente.

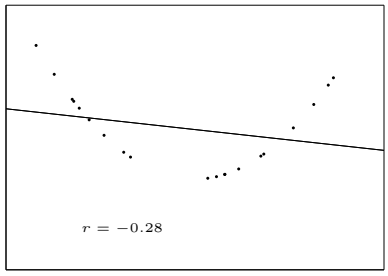


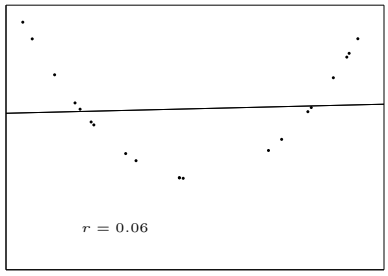
Nonostante il coefficiente di correlazione sia uguale a 0.96, il legame sembra sinusoidale piuttosto che lineare e quindi è più appropriato un modello nonlineare.



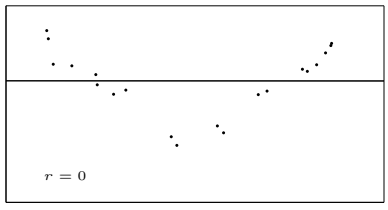


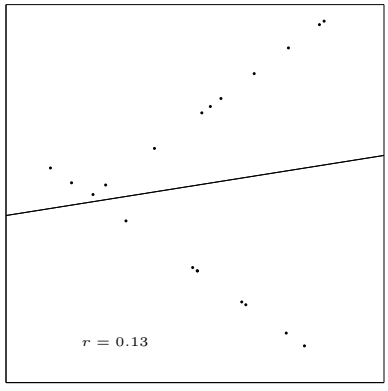
In questi due esempi il legame lineare è molto debole e nella seconda figura si ha l'impressione che la correlazione maggiore sia dovuta più a una certa simmetria e concentrazione al centro che a una dipendenza di y da x.



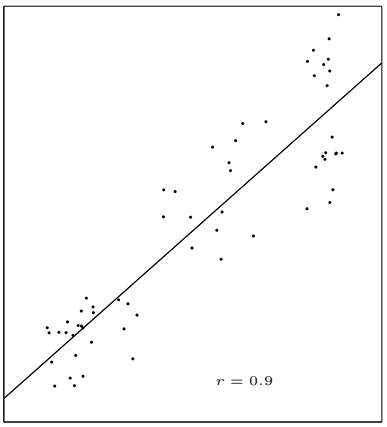


La dipendenza funzionale di tipo quadratico è evidente; il coefficiente di correlazione è vicino a 0; cfr. nota 8.14. Infatti il coefficiente di correlazione misura solo la dipendenza *lineare* tra le due variabili.

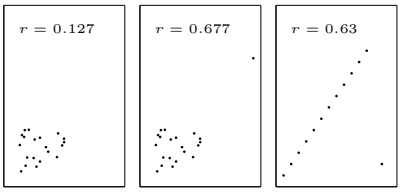




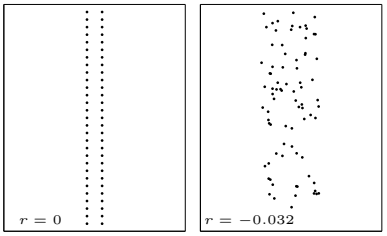
Nonostante che il coefficiente di correlazione sia molto vicino a zero, si notano in ciascuna delle ultime due figure due gruppi che esprimono una dipendenza lineare piuttosto spiccata di y da x. Questa situazione è tipica per dati non omogenei.



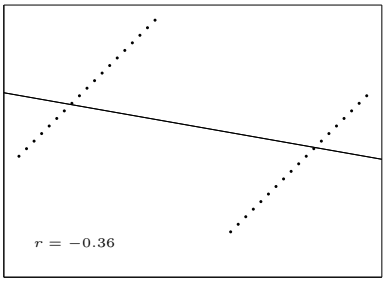
Anche questo è un caso di disomogeneità dei dati in cui però i tre gruppi distinti producono insieme un coefficiente di correlazione alto, benché all'interno di ogni gruppo la dipendenza lineare è piuttosto debole.



Si vede il forte effetto di un singolo valore eccezionale sul coefficiente di correlazione; persino nella seconda figura il coefficiente di correlazione è maggiore di quello nella terza!



Queste configurazioni illustrano un'altra volta quanto detto nella nota 8.14 riguardo al caso in cui i punti sono (almeno approssimativamente) simmetrici rispetto a una retta parallela all'asse delle y.



La correlazione totale è negativa, benché ogni gruppo presenti al suo interno una forte correlazione positiva.

Il quartetto di Anscombe

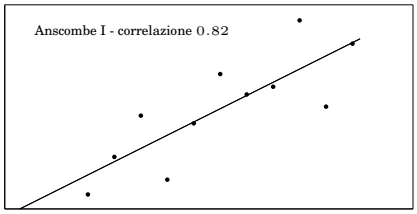
Esempi particolarmente impressivi sono stati costruiti da Francis Anscombe (citato in Bahrenberg/, 199-200). Consideriamo le seguenti serie di dati, noti nella letteratura come *quartetto di Anscombe*:

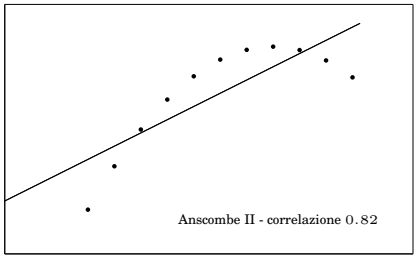
x_{I-III}	y_I	y_{II}	y_{III}	x_{IV}	y_{IV}
10.0	8.04	9.14	7.46	8.0	6.58
8.0	6.95	8.14	6.77	8.0	5.76
13.0	7.58	8.74	12.74	8.0	7.71
9.0	8.81	8.77	7.11	8.0	8.84
11.0	8.33	9.26	7.81	8.0	8.47
14.0	9.96	8.10	8.84	8.0	7.04
6.0	7.24	6.13	6.08	8.0	5.25
4.0	4.26	3.10	5.39	19.0	12.50
12.0	10.84	9.13	8.15	8.0	5.56
7.0	4.82	7.26	6.42	8.0	7.91
5.0	5.68	4.74	5.73	8.0	6.89

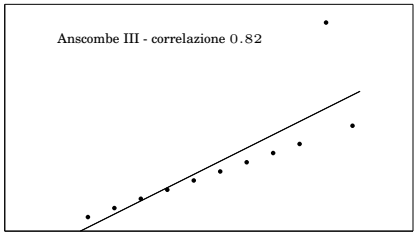
Questi dati hanno in comune le seguenti caratteristiche:

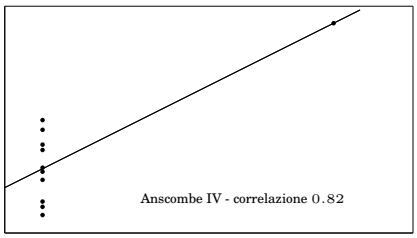
$n = 11$;
$\overline{x} = 9, \overline{y} = 7.5$;
retta di regressione $\eta = 0.5\xi + 0.3$;
coefficiente di correlazione $r = 0.82$.

Nonostante ciò le figure mostrano relazioni di dipendenza completamente diverse.









Solo nel primo caso l'analisi regressionale lineare può essere applicata. Gli esempi fanno vedere chiaramente che i valori numerici dei parametri statistici non sono sufficienti per una corretta interpretazione statistica che deve essere affiancata dalla rappresentazione grafica e uno studio il più dettagliato possibile dei meccanismi interni da cui i dati derivano.

Le critiche

Nel linguaggio comune il termine *correlazione* significa un rapporto stretto tra due elementi e questo significato viene spesso meccanicamente applicato al coefficiente di correlazione che invece deve essere compreso solo come un parametro numerico che non individua una precisa configurazione statistico-causale tra due variabili.

Il coefficiente di correlazione e i coefficienti della retta di regressione sono molto sensibili alla presenza anche di pochi valori eccezionali (in inglese *outliers*). Talvolta valori estremi possono essere semplicemente eliminati, ma ciò è permesso solo quando si può assumere che questi valori derivino da errori nelle misurazioni; in medicina valori estremi, quando non dovuti ad errori, hanno spesso significati diagnostici, per cui bisogna ricorrere ad un altro modello.

L'uso indiscriminato del coefficiente di correlazione viene spesso e giustamente criticato. J. Carroll chiama il coefficiente di correlazione

„one of the most frequently used tools of psychometricians ... and perhaps also one of the most frequently misused"

(citato in Rodgers/Nicewander, 61), e Arak Mathai, un famoso esperto di probabilità geometrica, è dell'opinione che il nome *coefficiente di correlazione* non dovrebbe essere più utilizzato, come risulta dalla recensione di uno dei suoi lavori sullo *Zentralblatt*:

„One of the most widely used concepts in statistical literature is the concept of correlation. In applied areas this correlation is interpreted as measuring relationship between variables. This article examines the structure of the expression defining correlation and shows that this concept cannot be meaningfully used to measure relationship or lack of it, or linearity or nonlinearity or independence or association or any such thing, and recommends that this misnomer correlation be replaced with something else in statistical literature."

„Il falsificatore astuto è più abile. Applica metodi formalmente inattaccabili a dati non adatti a questi metodi ..." (trad. da Fassl, 3)

Che nonostante le critiche, con un uso ragionato del coefficiente di correlazione si possono ottenere anche rappresentazioni molto convincenti di legami statistici, lo mostrano i grafici alle pagine 188-189 del libro di Bahrenberg/, in cui sono illustrate le correlazioni tra le diverse zone climatiche della Germania.

F. Anscombe: Graphs in statistical analysis. Am. Statistician 27 (1973), 17-21.

G. Bahrenberg/E. Giese/J. Nipper: Statistische Methoden in der Geographie I. Teubner 1999.

J. Carroll: The nature of the data, or how to choose a correlation coefficient. Am. Statistician 38 (1984), 58-60.

H. Fassl: Einführung in die medizinische Statistik. Barth 1999.

A. Mathai: The concept of correlation and misinterpretations. Int. J. Math. Stat. Sci. 7/2 (1998), ...

A. Mathai: On Pearson's statistic for goodness of fit. Int. J. Math. Stat. Sci. 7/2 (1998), ...

J. Rodgers/W. Nicewander: Thirteen ways to look at the correlation coefficient. Am. Statistician 42/1 (1988), 59-66.

Correlazione parziale

Talvolta una correlazione tra x ed y è riconducibile alla correlazione di entrambe le variabili con una terza variabile; per studiare questi influssi si introduce la *correlazione parziale*. Una breve discussione si trova a pagina 14 del corso di Statistica multivariata 2005/06.

III. IL TEOREMA SPETTRALE

Ortogonalità

Situazione 11.1. V sia uno spazio vettoriale reale di dimensione finita e $\| \ \|$ un prodotto scalare (cioè una forma bilineare simmetrica positivamente definita) su V.

Definizione 11.2. Due vettori $v, w \in V$ si chiamano *ortogonali* se $\|v, w\| = 0$. In questo caso scriviamo anche $v \perp w$.

Più in generale, per sottoinsiemi X, Y di V scriviamo $X \perp Y$ se $x \perp y$ per ogni $x \in X$ ed ogni $y \in Y$.

Definizione 11.3. X sia un sottoinsieme di V. Poniamo

$$X^\perp := \{v \in V \mid v \perp x \text{ per ogni } x \in X\}$$

Dalla bilinearità del prodotto scalare segue facilmente che $X^\perp$ è un sottospazio vettoriale di V (anche quando X stesso non è un sottospazio vettoriale).

Definizione 11.4. $W_1, \ldots, W_k$ siano sottospazi vettoriali di V. Diciamo che V è *somma ortogonale* di $W_1, \ldots, W_k$, se sono soddisfatte le seguenti condizioni:

(1) $V = W_1 + \ldots + W_k$.
(2) $W_i \perp W_j$ per $i \neq j$.

Scriviamo allora $V = W_1 \boxplus \ldots \boxplus W_k$.

Osservazione 11.5. *X ed Y siano sottoinsiemi di V tali che $X \perp Y$.*
Allora $X \cap Y \subset \{0\}$.
Se X ed Y sono sottospazi vettoriali, si ha quindi $X \cap Y = 0$.

Dimostrazione. Sia $v \in X \cap Y$. Per ipotesi allora $\|v, v\| = 0$ e ciò implica $v = 0$.

Corollario 11.6. *W sia un sottospazio vettoriale di V.*
Allora $W \cap W^\perp = 0$.

Osservazione 11.7. *$W_1, \ldots, W_k$ siano sottospazi vettoriali di V tali che $V = W_1 \boxplus \ldots \boxplus W_k$.*
Allora $V = W_1 \oplus \ldots \oplus W_k$, cioè $W_i \cap W_j = 0$ per $i \neq j$.

Dimostrazione. Osservazione 11.5.

Definizione 11.8. Per $v_1, \ldots, v_r \in V$ sia $\mathrm{SV}(v_1, \ldots, v_r)$ il sottospazio vettoriale generato da $v_1, \ldots, v_r$.

Osservazione 11.9. *I vettori $v_1, \ldots, v_r \in V$ siano ortogonali tra di loro e tutti $\neq 0$. Allora questi vettori sono anche linearmente indipendenti.*

Dimostrazione. Infatti sia $\alpha_1 v_1 + \ldots + \alpha_r v_r = 0$ per una scelta di coefficienti $\alpha_1, \ldots, \alpha_r \in \mathbb{R}$. Allora per ogni j abbiamo

$$0 = \|v_j, \alpha_1 v_1 + \ldots + \alpha_r v_r\| = \alpha_j \|v_j, v_j\|$$

Siccome $v_j \neq 0$ per ipotesi, segue $\alpha_j = 0$.

Nota 11.10 (ortonormalizzazione di Schmidt). $e_1, \ldots, e_s$ siano vettori linearmente indipendenti di V. Consideriamo vettori della forma

$$\begin{aligned} f_1 &:= e_1 \\ f_2 &:= e_2 - \alpha_{21} f_1 \\ f_3 &:= e_3 - \alpha_{31} f_1 - \alpha_{32} f_2 \\ &\ldots \\ f_s &:= e_s - \alpha_{s1} f_1 - \ldots - \alpha_{s,s-1} f_{s-1} \end{aligned}$$

con coefficienti reali α_{ij} che cerchiamo di determinare in modo tale che $f_k \perp f_j$ per $1 \leq j < k \leq s$.

Osserviamo in primo luogo che $\mathrm{SV}(f_1, \ldots, f_k) = \mathrm{SV}(e_1, \ldots, e_k)$ per ogni k. Ciò implica che $f_k \neq 0$ perché altrimenti, per $k \geq 2$, si avrebbe $e_k \in \mathrm{SV}(e_1, \ldots, e_{k-1})$ in contraddizione alla lineare indipendenza dei vettori e_j, mentre naturalmente anche $f_1 \neq 0$.

Sia $1 \leq j < k \leq s$. Le condizioni di ortogonalità che chiediamo significano

$$\|e_k - \alpha_{k1} f_1 - \ldots - \alpha_{k,k-1} f_{k-1}, f_j\| = 0$$

ovvero, usando per induzione che $f_i \perp f_j$ per $i < k$ ed $i \neq j$,

$$\|e_k, f_j\| - \alpha_{kj} \|f_j, f_j\| = 0, \text{ cosicché } \alpha_{kj} = \frac{\|e_k, f_j\|}{\|f_j, f_j\|}.$$

In questo modo abbiamo trovato un sistema $f_1, \ldots, f_s$ di vettori ortogonali tra di loro. Se poniamo $g_k := \dfrac{f_k}{|f_k|}$ per ogni k, otteniamo un sistema ortonormale. Siccome vettori $\neq 0$ ortogonali tra di loro sono linearmente indipendenti (osservazione 11.9), per $s = n$ troviamo in questo modo basi ortogonali risp. ortonormali di V.

Osservazione 11.11. *Se nella nota 11.10 i vettori $e_1, \ldots, e_s$ sono già ortogonali tra di loro, allora $f_k = e_k$ per ogni k.*

Dimostrazione. Infatti per la costruzione usata in questa ipotesi

$$e_k \perp \mathrm{SV}(e_1, \ldots, e_{k-1}) = \mathrm{SV}(f_1, \ldots, f_{k-1})$$

per $k \geq 2$ e ciò implica che $\alpha_{kj} = \dfrac{\|e_k, f_j\|}{\|f_j, f_j\|} = 0$ per ogni k, j.

Proposizione 11.12. *Ogni sottospazio vettoriale $W \neq 0$ di V possiede una base ortonormale ed ogni base ortonormale di W può essere estesa a una base ortonormale di V.*

Dimostrazione. Il primo enunciato segue dalla nota 11.10, il secondo dall'osservazione 11.11.

Teorema 11.13. *W sia un sottospazio vettoriale di V. Allora*

$$\dim V = \dim W + \dim W^\perp$$

Dimostrazione. Ciò è una conseguenza immediata della proposizione 11.12.

Lemma 11.14. *W sia un sottospazio vettoriale di V.*
Allora $W^{\perp\perp} = W$.

Dimostrazione. (1) Per la simmetria della relazione di ortogonalità è chiaro che ogni elemento di W è ortogonale ad ogni elemento di $W^\perp$, per cui $W \subset W^{\perp\perp}$.

(2) Per il teorema 11.13 abbiamo

$$\begin{aligned} \dim W^{\perp\perp} &= \dim V - \dim W^\perp \\ &= \dim V - (\dim V - \dim W) = \dim W \end{aligned}$$

$W \subset W^{\perp\perp}$ implica adesso $W = W^{\perp\perp}$.

Corollario 11.15. *W sia un sottospazio vettoriale di V. Allora*

$$W = V \iff W^\perp = 0$$

Dimostrazione. $\Longrightarrow$: Per il corollario 11.6 abbiamo $V \cap V^\perp = 0$. Ma $V \cap V^\perp = V^\perp$.

$\Longleftarrow$: Sia $W^\perp = 0$. Però $W \overset{11.14}{=} W^{\perp\perp} = 0^\perp = V$.

Lemma 11.16. *W_1 e W_2 siano sottospazi vettoriali di V. Allora*

$$\dim(W_1 + W_2) + \dim(W_1 \cap W_2) = \dim W_1 + \dim W_2$$

Dimostrazione. Corsi di Geometria.

Proposizione 11.17. *W sia un sottospazio vettoriale di V. Allora $V = W \boxplus W^\perp$.*

Dimostrazione. Siccome $W \cap W^\perp = 0$, dobbiamo solo dimostrare che $W + W^\perp = V$. Per il lemma 11.16 e il teorema 11.13

$$\begin{aligned} \dim(W + W^\perp) &= \dim(W + W^\perp) + \dim(W \cap W^\perp) \\ &= \dim W + \dim W^\perp = \dim V \end{aligned}$$

Lemma 11.18. *$e_1, \ldots, e_m$ sia una base ortonormale di V e $v, w \in V$ con*

$$\begin{aligned} v &= \alpha_1 e_1 + \ldots + \alpha_m e_m \\ w &= \beta_1 e_1 + \ldots + \beta_m e_m \end{aligned}$$

con $\alpha_i, \beta_j \in \mathbb{R}$. Allora $\|v, w\| = \sum_{k=1}^m \alpha_k \beta_k$.

Dimostrazione. Abbiamo

$$\begin{aligned} \|v, w\| &= \|\alpha_1 e_1 + \ldots + \alpha_m e_m, \beta_1 e_1 + \ldots + \beta_m e_m\| \\ &= \sum_{i=1}^m \sum_{j=1}^m \alpha_i \beta_j \|e_i, e_j\| = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \beta_j \delta_{ij} = \sum_{k=1}^m \alpha_k \beta_k \end{aligned}$$

Il teorema spettrale

Situazione 12.1. Come a pagina 11 sia V uno spazio vettoriale reale di dimensione finita e $\| \ \|$ un prodotto scalare su V. Supponiamo inoltre che $V \neq 0$ e che $\varphi : V \longrightarrow V$ sia un'applicazione lineare.

Definizione 12.2. φ si dice *simmetrica* se $\|\varphi v, w\| = \|v, \varphi w\|$ per ogni $v, w \in V$.

Definizione 12.3. Un *autovalore* di φ è un numero $\lambda \in \mathbb{C}$ tale che $\det(\varphi - \lambda \,\mathrm{id}) = 0$.

Se λ è un autovalore *reale* di φ, un *autovettore* di φ per λ è un elemento $v \in V \setminus 0$ per cui $\varphi v = \lambda v$.

Proposizione 12.4. *Se φ è simmetrica, allora ogni autovalore di φ è reale.*

Dimostrazione. Corsi di Geometria. Non è difficile, ma bisogna lavorare con spazi vettoriali su $\mathbb{C}$ e dimostrare l'enunciato analogo per operatori hermitiani.

Proposizione 12.5. *λ sia un autovalore reale di φ. Allora esiste un autovettore di λ in V.*

Dimostrazione. Siccome $\det(\varphi - \lambda \,\mathrm{id}) = 0$, l'applicazione

$$\varphi - \lambda \,\mathrm{id} : V \longrightarrow V$$

non è iniettiva e quindi $\mathrm{Ker}(\varphi - \lambda \,\mathrm{id}) \neq 0$; qui usiamo l'ipotesi che $V \neq 0$. Ma gli elementi di $\mathrm{Ker}(\varphi - \lambda \,\mathrm{id}) \setminus 0$ sono proprio gli autovettori di φ per λ.

Corollario 12.6. *Se φ è simmetrica, allora φ possiede un autovalore reale λ e un autovettore per λ.*

Dimostrazione. In primo luogo esiste un autovalore $\lambda \in \mathbb{C}$, perché $\mathbb{C}$ è algebricamente chiuso. L'enunciato segue dalle proposizioni 12.4 e 12.5.

Definizione 12.7. Per $\lambda \in \mathbb{R}$ sia

$$E_\lambda(\varphi) := \mathrm{Ker}(\varphi - \lambda \,\mathrm{id}) = \{v \in V \mid \varphi v = \lambda v\}$$

$E_\lambda(\varphi)$ è un sottospazio vettoriale di V che si chiama l'*autospazio* di φ rispetto al valore o autovalore λ.

Si noti che $E_\lambda(\varphi) \neq 0$ se e solo se λ è un autovalore di φ; ciò segue dalla proposizione 12.5 perché, per ipotesi, λ è reale.

È anche chiaro che $E_\lambda(\varphi)$ è φ-invariante: Se $\varphi v = \lambda v$, allora $\varphi\varphi v = \varphi(\lambda v) = \lambda \varphi v$.

Lemma 12.8. *φ sia simmetrica e $\lambda, \mu \in \mathbb{R}$ con $\lambda \neq \mu$.*
Allora $E_\lambda(\varphi) \perp E_\mu(\varphi)$.

Dimostrazione. Siano $v \in E_\lambda(\varphi)$ e $w \in E_\mu(\varphi)$.

Per l'ipotesi di simmetria $\|\varphi v, w\| = \|v, \varphi w\|$. Ma

$$\|\varphi v, w\| = \|\lambda v, w\| = \lambda \|v, w\| \quad \text{e} \quad \|v, \varphi w\| = \|v, \mu w\| = \mu \|v, w\|.$$

Siccome $\lambda \neq \mu$, necessariamente $\|v, w\| = 0$.

Corollario 12.9. *φ sia simmetrica e $\lambda, \mu \in \mathbb{R}$ con $\lambda \neq \mu$.*
Allora $E_\lambda(\varphi) \cap E_\mu(\varphi) = 0$.

Dimostrazione. Ciò segue dal lemma 12.8 e dall'osservazione 11.5.

Definizione 12.10. Un sottospazio vettoriale W di V si dice *φ-invariante*, se $\varphi W \subset W$.

In tal caso possiamo considerare l'applicazione lineare

$$\varphi_W := \bigcirc_w \varphi w : W \longrightarrow W$$

Proposizione 12.11. *φ sia simmetrica e W un sottospazio vettoriale φ-invariante di V. Allora anche $W^\perp$ è φ-invariante.*

Dimostrazione. Sia $v \in W^\perp$. Per ogni $w \in W$ abbiamo allora

$\|\varphi v, w\| = \|v, \varphi w\| = 0$, perché per ipotesi $\varphi w \in W$.

Osservazione 12.12. *W sia un sottospazio vettoriale di V. Allora la restrizione di $\| \ \|$ a W è un prodotto scalare su W.*

Osservazione 12.13. *W sia un sottospazio vettoriale φ-invariante di V. Se φ è simmetrica, anche φ_W è simmetrica.*

Corollario 12.14. *φ sia simmetrica e W un sottospazio vettoriale φ-invariante $\neq 0$ di V. Allora W contiene un autovettore di φ.*

Dimostrazione. Siccome $W \neq 0$, tenendo conto delle osservazioni 12.12 e 12.13 possiamo applicare il corollario 12.6 all'operatore simmetrico φ_W. È chiaro che un autovettore di φ_W è anche un autovettore di φ.

Teorema 12.15. *φ sia simmetrica e $\lambda_1, \ldots, \lambda_k$ gli autovalori distinti (necessariamente tutti reali) di φ. Allora*

$$V = E_{\lambda_1}(\varphi) \boxplus \ldots \boxplus E_{\lambda_k}(\varphi)$$

Dimostrazione. Sia $W := E_{\lambda_1}(\varphi) + \ldots + E_{\lambda_k}(\varphi)$.

Per il lemma 12.8 i sommandi sono ortogonali tra di loro. Dobbiamo quindi solo dimostrare che $W = V$. Per il corollario 11.15 è sufficiente dimostrare che $W^\perp = 0$.

Da quanto osservato alla fine della definizione 12.7 segue che W è φ-invariante, essendo somma di sottospazi φ-invarianti.

Sia $W^\perp \neq 0$. Dalla proposizione 12.11 sappiamo che $W^\perp$ è φ-invariante e dal corollario 12.14 segue che esistono $\mu \in \mathbb{R}$ e $v \in W^\perp \setminus 0$ tali che $\varphi v = \mu v$. Ma allora μ è un autovalore di φ, perciò esiste un j tale che $\mu = \lambda_j$. Ciò implica $v \in E_{\lambda_j}(\varphi)$ e quindi $v \in W$ perché, per costruzione, $E_{\lambda_j} \subset W$.

D'altra parte $v \in W^\perp$, quindi $v \in W \cap W^\perp$, e ciò, per il corollario 11.6, implica $v = 0$, una contraddizione.

Nota 12.16. φ sia simmetrica e $\lambda_1, \ldots, \lambda_k$ gli autovalori distinti di φ. Per il teorema 12.15

$$V = E_{\lambda_1}(\varphi) \boxplus \ldots \boxplus E_{\lambda_k}(\varphi)$$

Se per ogni j scegliamo in modo qualsiasi una base ortonormale di $E_{\lambda_j}(\varphi)$ (ciò è possibile per la proposizione 11.12), essa consiste necessariamente di autovettori di φ rispetto all'autovalore λ_j. Combinando tutte queste basi, otteniamo una base ortonormale di V consistente di autovettori di φ.

Corollario 12.17. *$A \in \mathbb{R}^s_s$ sia una matrice reale simmetrica. Allora esiste una matrice ortogonale U tale che $U^{-1}AU$ sia diagonale.*

Dimostrazione. Applichiamo la nota 12.16 al caso $V = \mathbb{R}^s$ con $\varphi := \bigcirc_x Ax$. È immediato che φ è simmetrica rispetto al prodotto scalare comune in $\mathbb{R}^s$. Per la nota 12.16 esiste una base ortonormale $e_1, \ldots, e_s$ che consiste di autovettori di A. Se U è la matrice le cui colonne sono gli e_j, otteniamo l'enunciato.

Decomposizione spettrale di operatori simmetrici

Nota 12.18. $W_1, \ldots, W_k$ siano sottospazi vettoriali di V tali che

$$V = W_1 \oplus \ldots \oplus W_k \quad (*)$$

Allora ogni $v \in V$ possiede un'unica rappresentazione nella forma

$$v = w_1 + \ldots + w_k$$

con $w_i \in W_i$ per ogni i. Se poniamo $\pi_i v := w_i$, otteniamo applicazioni $\pi_i : V \longrightarrow W_i$, che sono, come si dimostra facilmente, lineari e suriettive. Esse sono le *proiezioni* rispetto alla decomposizione (*). Per ogni $v \in V$ abbiamo $v = \pi_1 v + \ldots + \pi_k v$. Ciò corrisponde a una decomposizione

$$\mathrm{id} = \pi_1 + \ldots + \pi_k$$

dell'identità.

Nota 12.19. φ sia simmetrica e $\lambda_1, \ldots, \lambda_k$ gli autovalori distinti di φ. Applicando la nota 12.18 alla decomposizione

$$V = E_{\lambda_1}(\varphi) \boxplus \ldots \boxplus E_{\lambda_k}(\varphi)$$

abbiamo $\varphi \pi_i v = \lambda_i \pi_i v$ per ogni $v \in V$ e quindi

$$\varphi v = \varphi \pi_1 v + \ldots + \varphi \pi_k v = \lambda_1 \pi_1 v + \ldots + \lambda_k \pi_k v$$

ottenendo così la decomposizione spettrale

$$\varphi = \lambda_1 \pi_1 + \ldots + \lambda_k \pi_k$$

dell'operatore simmetrico φ.

Il rapporto di Rayleigh

Situazione 13.1. V sia come finora uno spazio vettoriale reale di dimensione finita $m \geq 1$ e $\| \ \|$ un prodotto scalare reale su V. $\varphi : V \longrightarrow V$ sia un'applicazione lineare *simmetrica* rispetto a $\| \ \|$, nel senso della definizione 12.2. $\lambda_1, \ldots, \lambda_m$ siano gli autovalori (necessariamente reali) di φ e $\lambda_1 \geq \ldots \geq \lambda_m$.

Definizione 13.2. Per $v \in V \setminus 0$ sia $\mathcal{R}v := \dfrac{\|v, \varphi v\|}{\|v, v\|}$ il *rapporto* (o quoziente) *di Rayleigh* di φ in v. Per un sottoinsieme $X \subset V$ sia

$$\mathcal{R}X := \{\mathcal{R}v \mid v \in X \setminus 0\}$$

$\mathcal{R}X$ si chiama l'*insieme di Rayleigh* di φ su X. $\mathcal{R}V$ nei libri di analisi numerica è chiamato spesso l'insieme dei valori di φ.

Il rapporto di Rayleigh è importante non soltanto in analisi numerica, ma anche in alcuni campi della matematica applicata: statistica, meccanica delle strutture, chimica quantistica.

Osservazione 13.3. *v sia un autovettore di φ per l'autovalore λ. Allora $\mathcal{R}v = \lambda$.*

Dimostrazione. Un autovettore è $\neq 0$, perciò il quoziente di Rayleigh è definito. Inoltre

$$\frac{\|v, \varphi v\|}{\|v, v\|} = \frac{\|v, \lambda v\|}{\|v, v\|} = \frac{\lambda \|v, v\|}{\|v, v\|} = \lambda$$

Osservazione 13.4. *Siano $v \in V \setminus 0$ ed $\alpha \in \mathbb{R} \setminus 0$. Allora $\mathcal{R}\alpha v = \mathcal{R}v$.*

In particolare $\mathcal{R}V = \mathcal{R}\{v \in V \mid |v| = 1\}$.

Proposizione 13.5. *$e_1, \ldots, e_m$ sia una base ortonormale di V e $v \in V$ con $v = \alpha_1 e_1 + \ldots + \alpha_m e_m$. Allora:*

(1) $\|v, v\| = \sum_{k=1}^{m} \alpha_k^2$.

(2) Se gli e_k sono autovettori di φ con $\varphi e_k = \lambda_k e_k$ per ogni k, allora

$$\|v, \varphi v\| = \sum_{k=1}^{m} \lambda_k \alpha_k^2$$

Dimostrazione. (1) segue dal lemma 11.18.

(2) L'ipotesi implica che

$$\begin{aligned} \varphi v &= \varphi(\alpha_1 e_1 + \ldots + \alpha_m e_m) = \alpha_1 \varphi e_1 + \ldots + \alpha_m \varphi e_m \\ &= \alpha_1 \lambda_1 e_1 + \ldots + \alpha_m \lambda_m e_m \end{aligned}$$

L'enunciato segue ancora dal lemma 11.18.

Lemma 13.6. *Siano dati numeri reali $a_1, \ldots, a_m$ con $\alpha := \min(a_1, \ldots, a_m)$, $\beta := \max(a_1, \ldots, a_m)$.*

Allora l'inviluppo convesso dell'insieme $\{a_1, \ldots, a_m\}$ è l'intervallo $[\alpha, \beta]$.

Dimostrazione. $I := [\alpha, \beta]$ è un insieme convesso che contiene tutti i punti dati. Dobbiamo dimostrare che I è il più piccolo insieme convesso con questa proprietà. Ma ciò è ovvio perché I è già l'inviluppo convesso del solo insieme $\{\alpha, \beta\}$.

Teorema 13.7. $\mathcal{R}V = [\lambda_m, \lambda_1]$.

Dimostrazione. Per la nota 12.16 esiste una base ortonormale $e_1, \ldots, e_m$ di V tale che $\varphi e_1 = \lambda_1 e_1, \ldots, \varphi e_m = \lambda_m e_m$.

(1) Sia $v \in V \setminus 0$, ad esempio $v = \alpha_1 e_1 + \ldots + \alpha_m e_m$. Per la proposizione 13.5 allora

$$\frac{\|v, \varphi v\|}{\|v, v\|} = \frac{\lambda_1 \alpha_1^2 + \ldots + \lambda_m \alpha_m^2}{\alpha_1^2 + \ldots + \alpha_m^2}$$

Per $1 \leq k \leq m$ sia $t_k := \dfrac{\alpha_k^2}{\alpha_1^2 + \ldots + \alpha_m^2}$.

Allora $t_k \geq 0$ e $t_1 + \ldots + t_m = 1$, e vediamo che

$$\frac{\|v, \varphi v\|}{\|v, v\|} = t_1 \lambda_1 + \ldots + t_m \lambda_m$$

appartiene all'inviluppo convesso dei numeri reali $\lambda_1, \ldots, \lambda_m$.

(2) Se viceversa sono dati numeri reali $t_1, \ldots, t_m \geq 0$ con $t_1 + \ldots + t_m = 1$, e se poniamo $\alpha_k := \sqrt{t_k}$ per ogni k, allora

$$\frac{\alpha_k^2}{\alpha_1^2 + \ldots + \alpha_m^2} = \frac{t_k}{t_1 + \ldots + t_m} = t_k$$

e, ponendo $v := \alpha_1 e_1 + \ldots + \alpha_m e_m$, come prima

$$t_1 \lambda_1 + \ldots + t_m \lambda_m = \frac{\|v, \varphi v\|}{\|v, v\|}$$

(3) Ciò mostra che $\mathcal{R}V$ coincide con l'inviluppo convesso dei numeri reali $\lambda_1, \ldots, \lambda_m$ e quindi, per il lemma 13.6, con $[\lambda_m, \lambda_1]$.

Corollario 13.8. $\lambda_1 = \max \mathcal{R}V, \quad \lambda_m = \min \mathcal{R}V$.

Calcolo matriciale

Nota 13.9. Per $A \in \mathbb{R}_m^n$ e $B \in \mathbb{R}_s^m$ abbiamo $AB \in \mathbb{R}_s^n$ con

$$(AB)_j^i = \sum_{\alpha=1}^{m} A_\alpha^i B_j^\alpha = A^i B_j$$

per ogni i, j.

Il prodotto matriciale fornisce un'applicazione $\mathbb{R}_m^n \times \mathbb{R}_s^m \longrightarrow \mathbb{R}_s^n$.

Corollario 13.10. *Per $A \in \mathbb{R}_n^p$ e $v \in \mathbb{R}^n$ si ha $Av \in \mathbb{R}^p$ con*

$$(Av)^i = \sum_{\alpha=1}^{n} A_\alpha^i v^\alpha = A^i v \quad \textit{per ogni } i.$$

Corollario 13.11. *Per $f \in \mathbb{R}_m$ e $B \in \mathbb{R}_s^m$ si ha $fB \in \mathbb{R}_s$ con*

$$(fB)_j = \sum_{\alpha=1}^{m} f_\alpha B_j^\alpha = fB_j \quad \textit{per ogni } j.$$

Corollario 13.12. *Per $A \in A_m^n$ e $B \in \mathbb{R}_s^m$ si hanno*

$$(AB)^i = A^i B = \sum_{\alpha=1}^{m} A_\alpha^i B^\alpha \quad \textit{e} \quad (AB)_j = AB_j = \sum_{\alpha=1}^{m} A_\alpha B_j^\alpha.$$

Dimostrazione. Per il corollario 13.11 abbiamo $(A^i B)_j = A^i B_j = (AB)_j^i$, e per il corollario 13.10 $(AB_j)^i = A^i B_j = (AB)_j^i$ per ogni i, j.

Corollario 13.13. *Siano $v \in \mathbb{R}^n$ ed $f \in \mathbb{R}_s$. Allora $vf \in \mathbb{R}_s^n$ con*

$$(vf)_j^i = v^i f_j \quad \textit{per ogni } i, j.$$

Dimostrazione. Ciò è un caso speciale della nota 13.9.

Proposizione 13.14. *Siano $A \in \mathbb{R}_m^n$ e $B \in \mathbb{R}_s^m$. Allora*

$$AB = \sum_{\alpha=1}^{m} A_\alpha B^\alpha$$

Dimostrazione. Per il corollario 13.13 abbiamo

$$(\sum_{\alpha=1}^{m} A_\alpha B^\alpha)_j^i = \sum_{\alpha=1}^{m} (A_\alpha B^\alpha)_j^i = \sum_{\alpha=1}^{m} A_\alpha^i B_j^\alpha = (AB)_j^i$$

Corollario 13.15. *Siano $A \in \mathbb{R}_m^n$ e $v \in \mathbb{R}^m$.*

Allora $Av = \sum_{\alpha=1}^{m} A_\alpha v^\alpha$.

Nota 13.16. Siano $A \in \mathbb{R}_m^n$ ed $f \in \mathbb{R}_m$. Allora

$$Af^t = \sum_{\alpha=1}^{m} A_\alpha f_\alpha$$

Af^t è quindi una combinazione lineare di $A_1, \ldots, A_m$ con i coefficienti $f_1, \ldots, f_m$.

Useremo questa osservazione fra poco per $A = CX$.

Osservazione 13.17. *Sia $A \in \mathbb{R}_m^n$. Allora*

$$(A^i)^t = (A^t)_i \quad \textit{e} \quad (A_j)^t = (A^t)^j$$

per ogni i, j.

Spazi ortogonali intermedi

Proposizione 14.1. *Per ogni $v \in V$ si ha*

$$\lambda_m \|v, v\| \leq \|v, \varphi v\| \leq \lambda_1 \|v, v\|$$

Dimostrazione. Ciò per $v \neq 0$ segue dal teorema 13.7.

Lemma 14.2. *$e_1, \ldots, e_m$ sia una base ortonormale di V tale che $\varphi e_1 = \lambda_1 e_1, \ldots, \varphi e_m = \lambda_m e_m$. Sappiamo dalla nota 12.16 che una tale base esiste. Per $1 \leq r \leq s \leq m$ poniamo $E_{rs} := \mathbb{R}e_r + \ldots + \mathbb{R}e_s$.*

Allora $\mathcal{R}E_{rs} = [\lambda_s, \lambda_r]$ e quindi

$$\max \mathcal{R}E_{rs} = \lambda_r = \mathcal{R}e_r$$
$$\min \mathcal{R}E_{rs} = \lambda_s = \mathcal{R}e_s$$

Dimostrazione. È chiaro che E_{rs} è φ-invariante e che la matrice di $\psi := \bigcirc_v \varphi v : E_{rs} \longrightarrow E_{rs}$ rispetto alla base $e_r, \ldots, e_s$ è

$$\begin{pmatrix} \lambda_r & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & \lambda_s \end{pmatrix}$$

Ciò mostra che gli autovalori di ψ sono $\lambda_r, \ldots, \lambda_s$. ψ è simmetrica per l'osservazione 12.13 e soddisfa le condizioni della situazione 13.1; possiamo quindi applicare il teorema 13.7 a ψ.

Siccome $\lambda_r = \max(\lambda_r, \ldots, \lambda_s)$ e $\lambda_s = \min(\lambda_r, \ldots, \lambda_s)$, otteniamo l'enunciato, tenendo conto dell'uguaglianza $\lambda_i = \mathcal{R}e_i$ per ogni i che segue dall'osservazione 13.3.

Corollario 14.3. *$e_1, \ldots, e_m$ sia una base ortonormale di V tale che $\varphi e_1 = \lambda_1 e_1, \ldots, \varphi e_m = \lambda_m e_m$. Usiamo le notazioni del lemma 14.2.*

(1) Sia $1 \leq k \leq m$. Allora
$$\max \mathcal{R}E_{km} = \min \mathcal{R}E_{1k} = \lambda_k = \mathcal{R}e_k$$

(2) Sia $2 \leq k \leq m$. Allora
$$\max \mathcal{R}E_{1,k-1}^{\perp} = \max \mathcal{R}E_{km} = \lambda_k = \mathcal{R}e_k$$

(3) Sia $1 \leq k \leq m-1$. Allora
$$\min \mathcal{R}E_{k+1,m}^{\perp} = \min \mathcal{R}E_{1k} = \lambda_k = \mathcal{R}e_k$$

Osservazione 14.4. *X ed Y siano sottoinsiemi di V e $v \in X \cap Y$ con $v \neq 0$. Allora $\min \mathcal{R}X \leq \mathcal{R}v \leq \max \mathcal{R}Y$.*

Lemma 14.5. *W_1 e W_2 siano sottospazi vettoriali di V. Allora*

$$m + \dim(W_1 \cap W_2) \geq \dim W_1 + \dim W_2$$

Dimostrazione. Per il lemma 11.16

$$\begin{aligned} \dim W_1 + \dim W_2 &= \dim(W_1 + W_2) + \dim(W_1 \cap W_2) \\ &\leq m + \dim(W_1 \cap W_2) \end{aligned}$$

Lemma 14.6. *Sia $1 \leq k \leq m-1$. W sia un sottospazio vettoriale di V con $\dim W \leq k$. Allora $\max \mathcal{R}W^{\perp} \geq \lambda_{k+1}$.*

Dimostrazione. Scegliamo di nuovo una base ortonormale $e_1, \ldots, e_m$ di V tale che $\varphi e_1 = \lambda_1 e_1, \ldots, \varphi e_m = \lambda_m e_m$. Usiamo la notazione del lemma 14.2.

Per la proposizione 11.17 abbiamo $V = W \oplus W^{\perp}$ e quindi $\dim W^{\perp} = m - \dim W$. Per il lemma 14.5 abbiamo

$$\begin{aligned} m + \dim(W^{\perp} \cap E_{1,k+1}) &\geq \dim W^{\perp} + \dim E_{1,k+1} \\ &= m - \dim W + k + 1 \geq m - k + k + 1 = m + 1 \end{aligned}$$

Ciò implica $\dim(W^{\perp} \cap E_{1,k+1}) \geq 1$, per cui $W^{\perp} \cap E_{1,k+1} \neq 0$. Esiste quindi un vettore $v \in W^{\perp} \cap E_{1,k+1}$ con $v \neq 0$. Con il lemma 14.2 e usando l'osservazione 14.4 segue adesso $\lambda_{k+1} = \min \mathcal{R}E_{1,k+1} \leq \mathcal{R}v \leq \max \mathcal{R}W^{\perp}$.

Teorema 14.7 (teorema di Courant). *Sia $1 \leq k \leq m-1$. Sia $\mathcal{U}$ l'insieme dei sottospazi vettoriali U di V con $\dim U \geq m - k$. Allora*

$$\lambda_{k+1} = \min \{\max \mathcal{R}U \mid U \in \mathcal{U}\}$$

Dimostrazione. Sia $\mathcal{U}'$ l'insieme dei sottospazi vettoriali W di V con $\dim W \leq k$. Per il corollario 14.3 e il lemma 14.6 allora

$$\lambda_{k+1} = \min \{\max \mathcal{R}W^{\perp} \mid W \in \mathcal{U}'\}$$

Siccome $\dim W \leq k$ se e solo se $\dim W^{\perp} \geq m - k$, l'enunciato segue dalla proposizione 11.17.

Matrici normali

Una matrice $A \in \mathbb{R}_m^m$ si dice *normale*, se $AA^t = A^tA$. Una matrice simmetrica è evidentemente normale, ma anche ogni matrice antisimmetrica (cioè tale che $A^t = -A$) e ogni matrice ortogonale (cioè tale che $A^t = A^{-1}$) è normale. Matrici antisimmetriche o ortogonali non hanno in genere autovalori reali, si può però dimostrare che, se con $\varphi := \bigcirc_x Ax$ definiamo $\mathcal{R}V$ come nella definizione 25.4 (rispetto al prodotto scalare comune), $\mathcal{R}V$ coincide anche in questo caso con l'inviluppo convesso dell'insieme degli autovalori di A.
Cfr. Stoer/Bulirsch, pag. 85.

D. Bini/M. Capovani/O. Menchi: Metodi numerici per l'algebra lineare. Zanichelli 1988.

F. Paset: Regressione, correlazione e analisi delle componenti principali. Tesi Univ. Ferrara 2003.

J. Stoer/R. Bulirsch: Einführung in die numerische Mathematik II. Springer 1978.

Formule per il prodotto scalare

Nota 14.8. Con $\| \ \|$ denotiamo, come finora, il prodotto scalare comune sia in $\mathbb{R}^n$ che in $\mathbb{R}_m$.

Per $u, v \in \mathbb{R}^n$ abbiamo quindi $\|u, v\| = u^t v = \sum_{\alpha=1}^{m} u^\alpha v^\alpha$, mentre per $f, g \in \mathbb{R}_m$ abbiamo $\|f, g\| = fg^t = \sum_{\alpha=1}^{m} f_\alpha g_\alpha$.

Osservazione 14.9. *Per $A \in \mathbb{R}_n^n$ e $v, w \in \mathbb{R}^n$ risp. $B \in \mathbb{R}_m^m$ ed $f, g \in \mathbb{R}_m$ valgono*

$$v^t Aw = \|v, Aw\| = \|A^t v, w\|$$
$$fBg^t = \|fB, g\| = \|f, gB^t\|$$

Dimostrazione. $\|v, Aw\| = v^t Aw = (A^t v)^t w = \|A^t v, w\|$,

$\|fB, g\| = fBg^t = f(gB^t)^t = \|f, gB^t\|$.

Corollario 14.10. *Siano $A \in \mathbb{R}_p^n, B \in \mathbb{R}_s^n$.*
Allora $(A^tB)_j^i = \|A_i, B_j\|$ per ogni i, j.

Dimostrazione. Usando la nota 13.9 e l'osservazione 13.17 abbiamo

$$(A^tB)_j^i = (A^t)^i B_j = (A_i)^t B_j = \|A_i, B_j\|.$$

Lemma 14.11. *Per $A \in \mathbb{R}_p^n$ e $v \in \mathbb{R}^n$ abbiamo $\|A_j, v\| = (A^t v)^j$ per ogni j.*

Dimostrazione. Dal corollario 13.12 e dall'osservazione 13.17 segue $\|A_j, v\| = (A_j)^t v = (A^t)^j v = (A^t v)^j$.

Proposizione 14.12. *Siano $A \in \mathbb{R}_p^n$ e $v \in \mathbb{R}^n$. Allora*

$$\sum_{j=1}^{p} \|A_j, v\|^2 = \|v, AA^t v\| = v^t AA^t v$$

Dimostrazione. Dal lemma 14.11 e dall'osservazione 14.9 abbiamo

$$\sum_{j=1}^{p} \|A_j, v\|^2 = \sum_{j=1}^{p} ((A^t v)^j)^2 = \|A^t v, A^t v\| = \|v, AA^t v\|.$$

Lemma 14.13. *Per $f \in \mathbb{R}_m$ e $B \in \mathbb{R}_m^n$ abbiamo $\|f, B^i\| = (fB^t)_i$ per ogni i.*

Dimostrazione. Dal corollario 13.12 e dall'osservazione 13.17 segue

$$\|f, B^i\| = f(B^i)^t = f(B^t)_i = (fB^t)_i$$

Proposizione 14.14. *Siano $f \in \mathbb{R}_m$ e $B \in \mathbb{R}_m^n$. Allora*

$$\sum_{i=1}^{n} \|f, B^i\|^2 = \|fB^tB, f\| = fB^tBf^t$$

Dimostrazione. Usando il lemma 14.13 e l'osservazione 14.9 abbiamo

$$\sum_{i=1}^{n} \|f, B^i\|^2 = \sum_{i=1}^{n} ((fB^t)_i)^2 = \|fB^t, fB^t\| = \|fB^tB, f\|$$

La matrice A^tA

Osservazione 15.1. *Sia $A \in \mathbb{R}^n_m$. Allora le colonne di A sono linearmente dipendenti se e solo se esiste $x \in \mathbb{R}^m \setminus 0$ tale che $Ax = 0$.*

Dimostrazione. $Ax = \sum_{k=1}^{m} A_k x^k$ è una combinazione lineare delle colonne di A e ogni tale combinazione lineare può essere scritta in questo modo.

Osservazione 15.2. *Sia $A \in \mathbb{R}^n_m$. Allora la matrice $A^tA \in \mathbb{R}^m_m$ è simmetrica. Inoltre:*

(1) A^tA è positivamente semidefinita.

(2) A^tA è positivamente definita se e solo se le colonne di A sono linearmente indipendenti.

Dimostrazione. (1) Sia $x \in \mathbb{R}_m$. Allora

$$x^tA^tAx = (Ax)^tAx = \|Ax, Ax\| \geq 0.$$

(2) A^tA sia positivamente definita ed $x \in \mathbb{R}^m$ tale che $Ax = 0$. Ciò implica $x^tA^tAx = 0$ e quindi $x = 0$.

Siano viceversa le colonne di A linearmente indipendenti. Sia $x^tA^tAx = 0$, cioè $\|Ax, Ax\| = 0$. Allora $Ax = 0$ e quindi $x = 0$.

Corollario 15.3. *Sia $A \in \mathbb{R}^n_m$. Allora gli autovalori di A^tA sono ≥ 0. Essi sono tutti > 0 se e solo se le colonne di A sono linearmente indipendenti.*

La traccia

Definizione 15.4. Sia $A \in \mathbb{R}^m_m$. Definiamo la *traccia* di A, denotata con $\operatorname{tr} A$, come la somma degli elementi della diagonale principale di A, quindi $\operatorname{tr} A := \sum_{i=1}^{m} A^i_i$.

È chiaro che l'applicazione $\operatorname{tr} : \mathbb{R}^m_m \longrightarrow \mathbb{R}$ è lineare. La traccia gode però di molte altre proprietà importanti, tra cui i sorprendenti corollari 15.6 e 15.7.

Proposizione 15.5. *Siano $A \in \mathbb{R}^n_m$ e $B \in \mathbb{R}^m_n$. Allora $\operatorname{tr} AB = \operatorname{tr} BA$. Si noti che $AB \in \mathbb{R}^n_n$, mentre $BA \in \mathbb{R}^m_m$.*

Dimostrazione. Abbiamo

$$\operatorname{tr} AB = \sum_{i=1}^{n} (AB)^i_i = \sum_{i=1}^{n}\sum_{j=1}^{m} A^i_j B^j_i = \sum_{j=1}^{m}\sum_{i=1}^{n} B^j_i A^i_j$$
$$= \sum_{j=1}^{m} (BA)^j_j = \operatorname{tr} BA$$

Corollario 15.6. *Siano $f \in \mathbb{R}_m$ e $v \in \mathbb{R}^m$. Allora $fv = \operatorname{tr} vf$.*

Corollario 15.7. *(1) Siano $A \in \mathbb{R}^n_p$ e $v \in \mathbb{R}^n$. Allora*

$$\sum_{j=1}^{p} \|A_j, v\|^2 = v^tAA^tv = \operatorname{tr} A^tvv^tA$$

(2) Siano $f \in \mathbb{R}_m$ e $B \in \mathbb{R}^n_m$. Allora

$$\sum_{i=1}^{n} \|f, B^i\|^2 = fB^tBf^t = \operatorname{tr} Bf^tfB^t$$

Dimostrazione. Per le proposizioni 14.12 e 14.14 abbiamo $\sum_{j=1}^{p} \|A_j, v\|^2 = v^tAA^tv$ e $\sum_{i=1}^{n} \|f, B^i\|^2 = fB^tBf^t$. L'enunciato segue dalla proposizione 15.5.

Nota 15.8. Sia $A \in \mathbb{R}^n_m$. Allora

$$\operatorname{tr} A^tA = \operatorname{tr} AA^t = \sum_{i=1}^{n} |A^i|^2 = \sum_{j=1}^{m} |A_j|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} (A^i_j)^2$$

non è altro che il quadrato della lunghezza di A considerata come vettore di $\mathbb{R}^{mn}$.

Inversione al cerchio unitario

Nota 15.9. Chiamiamo *inversione al cerchio unitario* l'applicazione $K : \mathbb{C} \setminus 0 \longrightarrow \mathbb{C}$ che manda ogni punto $z \neq 0$ nel punto z' che si trova sulla semiretta che parte dall'origine e passa per z, avendo però un modulo che è il reciproco di quello di z, cioè tale che $|z'| = \dfrac{1}{|z|}$.

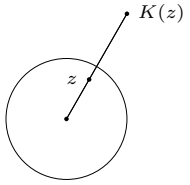
Questa applicazione è talvolta anche detta *riflessione* al cerchio unitario. K è univocamente determinata dalla condizione enunciata e

$$K(z) = \frac{z}{|z|^2} = \frac{1}{\overline{z}} \text{ per ogni } z \in \mathbb{C} \setminus 0.$$

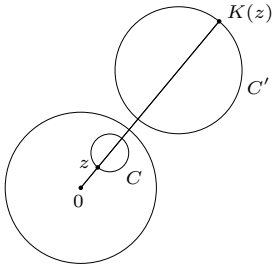
Dimostrazione. Infatti dobbiamo avere $K(z) = tz$ con $t > 0$ e inoltre deve valere $|tz| = \dfrac{1}{|z|}$; essendo $t > 0$ ciò è equivalente a $t|z| = \dfrac{1}{|z|}$, cioè $t = \dfrac{1}{|z|^2}$. Quindi l'immagine $K(z) = \dfrac{z}{|z|^2}$ è univocamente determinata.

Osservando che $|z|^2 = z\overline{z}$, vediamo che $K(z) = \dfrac{z}{z\overline{z}} = \dfrac{1}{\overline{z}}$.

Esercizio 15.10. Verificare da soli che $K \circ K = \mathrm{id}$; è invece chiaro direttamente dalla definizione che i punti fissi di K sono esattamente i punti del cerchio unitario e che ogni punto all'interno del cerchio unitario viene trasformato in un punto esterno e viceversa.

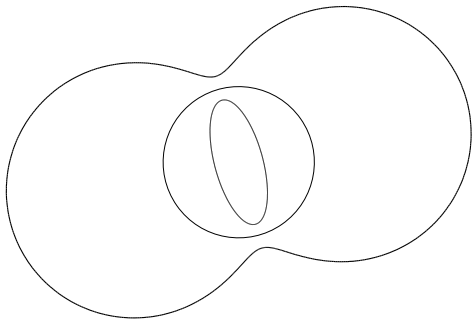


L'inversione al cerchio unitario possiede interessanti proprietà geometriche, molte delle quali sono descritte nel bellissimo libro di Needham e di cui la più importante è quella che attraverso K ogni cerchio C che non passa per l'origine viene trasformato in un cerchio C' (anch'esso non passante per l'origine).



Si noti che il centro di C' non è l'immagine del centro di C!

Osservazione 15.11. Vedremo adesso che, mentre l'inversione al cerchio unitario trasforma cerchi in cerchi, un'ellisse (che non sia un cerchio) con centro nell'origine viene invece trasformata in una curva di quarto grado (una lemniscata ellittica); infatti la geometria dell'ellisse è molto più profonda e difficile della geometria del cerchio.



La lemniscata ellittica

Nota 16.1. Consideriamo un'ellisse nel piano con centro nell'origine, descritta da un'equazione

$$f(x,y) = 1$$

dove $f(x,y) = \alpha x^2 + 2\beta xy + \gamma y^2$ è una forma quadratica (reale) positivamente definita. Per $z = x + iy$ scriviamo anche $f(z)$ invece di $f(x,y)$. Per ogni $t \in \mathbb{R}$ allora $f(tz) = t^2 f(z)$. È ovvio inoltre che l'origine non appartiene all'ellisse, perché $f(0,0) = 0 \neq 1$. Denotiamo di nuovo con K l'inversione al cerchio unitario.

Siano adesso z un punto dell'ellisse e

$$w := K(z) = \frac{z}{|z|^2}$$

Allora

$$f(w) = \frac{1}{|z|^4} f(z) = \frac{1}{|z|^4}$$

perché $f(z) = 1$ essendo z un punto dell'ellisse. D'altra parte abbiamo

$$|z| = \frac{1}{|w|}$$

per definizione di K e quindi $f(w) = |w|^4$. I punti w dell'immagine dell'ellisse sotto K soddisfano quindi l'equazione

$$|w|^4 = f(w)$$

che, se poniamo $w = u + iv$ con $u, v \in \mathbb{R}$, può essere scritta anche nella forma

$$(u^2 + v^2)^2 = f(u,v)$$

oppure, ancora più esplicitamente,

$$(u^2 + v^2)^2 = \alpha u^2 + 2\beta uv + \gamma v^2$$

Si osservi però che oltre ai punti riflessi dell'ellisse anche l'origine soddisfa questa equazione. Curve con questa equazione si chiamano *lemniscate ellittiche* quando, come nella nostra ipotesi, la forma quadratica f è positivamente definita.

Nota 16.2. Vediamo adesso che la lemniscata ellittica può essere utilizzata per rappresentare il quoziente di Rayleigh di una forma quadratica in due dimensioni.

Nelle ipotesi e con la notazione della nota 16.1 sia z_0 il punto sulla circonferenza unitaria determinato da z, cioè $z_0 = \frac{z}{|z|}$.

Siano

$$L := \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$$

e $\varphi := \bigcirc_p pL : \mathbb{R}_2 \longrightarrow \mathbb{R}_2$ l'operatore simmetrico associato.

Allora $\|p, \varphi p\| = f(p)$ è vediamo che $f(z_0) = \mathcal{R} z_0$ è proprio il quoziente di Rayleigh di φ in z_0. D'altra parte

$$f(z_0) = \frac{1}{|z|^2} f(z) = \frac{1}{|z|^2}$$

e quindi $\frac{1}{|z|} = \sqrt{f(z_0)}$ oppure, equivalentemente,

$$|z| = \frac{1}{\sqrt{f(z_0)}}$$

Questa prima equazione mostra che il modulo di un punto z dell'ellisse $f(z) = 1$ è uguale a $\frac{1}{\sqrt{f(z_0)}}$, dove z_0 è il vettore unitario che mostra nella stessa direzione di z.

Oltre a ciò abbiamo però anche

$$w = \frac{z}{|z|^2} = \frac{z_0}{|z|} = z_0 \cdot \sqrt{f(z_0)}$$

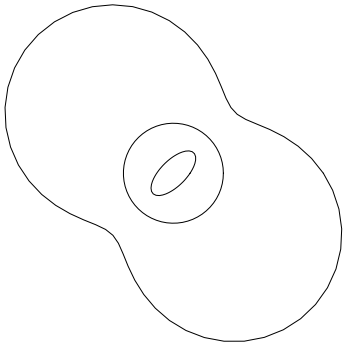
quindi

$$w = z_0 \cdot \sqrt{f(z_0)}$$

Osservazione 16.3. Il punto w che si ottiene da un punto z dell'ellisse mediante inversione al cerchio unitario è quindi quel vettore che si ottiene moltiplicando il vettore unitario z_0 con la radice $\sqrt{f(z_0)}$ del quoziente di Rayleigh in quella direzione.

Nel caso statistico del capitolo 4 di $L = \Omega$ con $m = 2$ il fattore $\sqrt{f(z_0)}$ è uguale alla deviazione standard di X_{z_0}.

Presentiamo una realizzazione in R, perché è semplicissima.



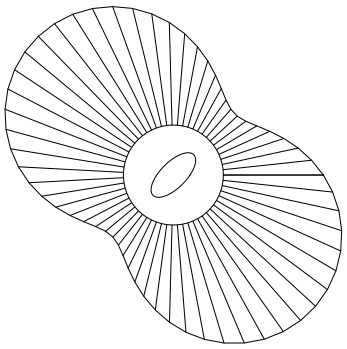
Otteniamo questa figura, che corrisponde alla quadrica $f(x,y) = 9x^2 - 12xy + 9y^2 = 1$, con il programma

```
f = function (x,y)
9*x^2-12*x*y+9*y^2

# Radice di f(z).
rfz = function (z)
{x=Re(z); y=Im(z)
sqrt(f(x,y))}

Gr.postscript('27-lem-1.ps',5,5)
larg=7
intervallo=c(-larg/2,larg/2)
Gr(intervallo,intervallo)
alfa=seq(0,2*pi,length=60)
z=exp(1i*alfa)
lines(z)
lines(z/rfz(z))
lines(z*rfz(z))
dev.off()
```

Le funzioni `Gr.postscript` e `Gr` sono semplici funzioni di impostazione grafica. Aggiungendo le righe

```
raggi=Mm(c(z,z*rfz(z),
  rep(NA,length(z))),righe=3)
lines(raggi)
```

possiamo evidenziare i raggi tra i punti z_0 del cerchio unitario e i corrispondenti punti w sulla lemniscata.



A. Coffman/M. Frantz: Möbius transformations and ellipses. Internet 2004, 9p.

T. Needham: Visual complex analysis. Oldenbourg 2001.

IV. ANALISI DELLE COMPONENTI PRINCIPALI

Le matrici MX e CX

Situazione 17.1. Sia $X \in \mathbb{R}^n_m$ con $n \geq 2$.
Usiamo le notazioni della def. 1.1. A pagina 2 abbiamo introdotto le matrici M e $C = \delta - M$.

Osservazione 17.2. $MX = (\overline{X_1}^\diamond, \ldots, \overline{X_m}^\diamond)$. *Più esplicitamente*

$$MX = \begin{pmatrix} \overline{X_1} & \ldots & \overline{X_m} \\ \vdots & & \vdots \\ \overline{X_1} & \ldots & \overline{X_m} \end{pmatrix}$$

Dimostrazione. Per definizione $X = (X_1, \ldots, X_m)$.

per cui $MX = (MX_1, \ldots, MX_m)$.

L'enunciato segue dall'osservazione 2.4.

Nota 17.3. Nel caso $m = 2$ abbiamo perciò

$$MX = M(x, y) = (\overline{x}^\diamond, \overline{y}^\diamond) = \begin{pmatrix} \overline{x} & \overline{y} \\ \vdots & \vdots \\ \overline{x} & \overline{y} \end{pmatrix}$$

Definizione 17.4. $\overline{X} := (\overline{X_1}, \ldots, \overline{X_m})$ è il baricentro delle righe di X. Si noti che $\overline{X} \in \mathbb{R}_m$, mentre $MX \in \mathbb{R}^n_m$.

Definizione 17.5. La matrice $CX = X - MX$ si chiama la matrice dei dati *centralizzata*. Come nell'osservazione 17.2

$$CX = (CX_1, \ldots, CX_m) = (X_1 - \overline{X_1}^\diamond, \ldots, X_m - \overline{X_m}^\diamond)$$

Nel caso $m = 2$ abbiamo $C(x, y) = (Cx, Cy)$.

Nella letteratura la matrice $X - MX$ viene talvolta anche chiamata la *matrice delle deviazioni* (dalle medie).

Il baricentro

Lemma 17.6. *$f^1, \ldots, f^n$ siano punti in $\mathbb{R}_m$ e $b = \dfrac{f^1 + \ldots + f^n}{n}$ il loro baricentro. Per $c \in \mathbb{R}_m$ sia $F(c) := \sum_{i=1}^n |f^i - c|^2$.*

Allora $F(b) < F(c)$ per ogni $c \neq b$.

Dimostrazione. Con $c := (c_1, \ldots, c_m)$ abbiamo

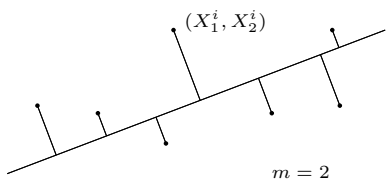
$$F(c_1, \ldots, c_m) = \sum_{i=1}^n \sum_{k=1}^m (f^i_k - c_k)^2$$

Per la determinazione del minimo consideriamo le equazioni

$\dfrac{\partial F}{\partial c_j} = -2\sum_{i=1}^n (f^i_j - c_j) = 0$, cioè $nc_j = \sum_{i=1}^n f^i_j$ e ciò implica $c = \dfrac{1}{n}(f^1 + \ldots + f^n)$. Siccome $\dfrac{\partial F}{\partial c_j \partial c_k} = 2\delta_{jk}$ per ogni j, k, vediamo che si tratta di un minimo che deve essere un minimo assoluto; infatti F tende all'infinito per $c \longrightarrow \infty$, per cui possiamo limitarci a cercare il minimo in un disco compatto attorno all'origine.

Regressione ortogonale

Nota 17.7. Nella *regressione ortogonale* cerchiamo una retta in $\mathbb{R}_m$ tale da minimizzare la somma dei quadrati delle distanze dei punti X^i da questa retta; cfr. pagina 5. Ogni retta con questa proprietà si chiama una *retta di regressione ortogonale* di X.

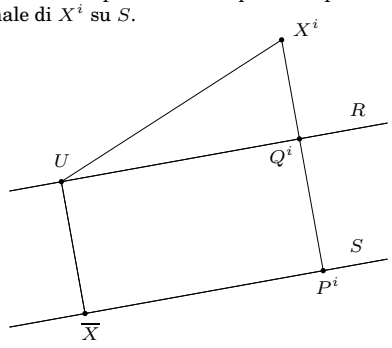


Osservazione 17.8. *Esiste almeno una retta di regressione ortogonale di X.*

Dimostrazione. Infatti la funzione da minimizzare è ovviamente continua ed è evidente che ci si può limitare a un insieme compatto di parametri da variare.

Lemma 17.9. *Ogni retta di regressione ortogonale passa per il baricentro $\overline{X} = (\overline{X_1}, \ldots, \overline{X_m})$ dei punti X^i.*

Dimostrazione. R sia una retta di regressione ortogonale. U sia la proiezione ortogonale di $\overline{X}$ su R. Assumiamo che $\overline{X} \notin R$. Allora $U \neq \overline{X}$. Per ogni i sia Q^i la proiezione ortogonale di X^i su R.
Siano S la retta parallela ad R passante per $\overline{X}$ e P^i la proiezione ortogonale di X^i su S.



Allora $Q^i - U = P^i - \overline{X}$ per ogni i. Dal teorema di Pitagora abbiamo

$$|X^i - Q^i|^2 = |X^i - U|^2 - |Q^i - U|^2$$
$$|X^i - P^i|^2 = |X^i - \overline{X}|^2 - |P^i - \overline{X}|^2 = |X^i - \overline{X}|^2 - |Q^i - U|^2$$

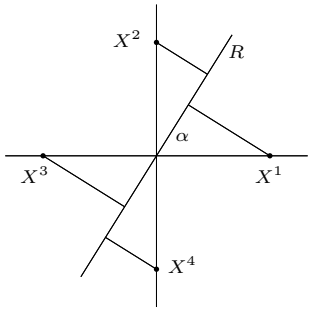
Dal lemma 17.6 segue adesso, contrariamente all'ipotesi,

$$\sum_{i=1}^n |X^i - P^i|^2 = \sum_{i=1}^n |X^i - \overline{X}|^2 - \sum_{i=1}^n |Q^i - U|^2$$
$$< \sum_{i=1}^n |X^i - U|^2 - \sum_{i=1}^n |Q^i - U|^2 = \sum_{i=1}^n |X^i - Q^i|^2$$

Nota 17.10. La retta di regressione ortogonale in genere non è univocamente determinata. Consideriamo ad esempio i quattro punti

$$X^1 = (1, 0), X^2 = (0, 1), X^3 = (-1, 0), X^4 = (0, -1)$$

nel piano $\mathbb{R}_2$. Per il lemma 17.9 ogni retta di regressione ortogonale R passa per il baricentro che in questo caso coincide con l'origine. Come nella figura sia α l'angolo tra R e l'ascisse.



Dal disegno si vede che la somma dei quadrati delle distanze dei punti X^i dalla retta R è uguale a $2(\sin^2\alpha + \cos^2\alpha) = 2$, indipendentemente da α.

In questo esempio entrambe le colonne di X hanno media 0 e ciò implica $MX = 0$. La matrice Ω che verrà definita a pagina 18 è quindi, per il lemma 18.3, uguale a

$$X^t X = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2\delta$$

Ciò significa che ogni vettore $\neq 0$ di $\mathbb{R}_2$ è un autovettore di Ω per l'unico autovalore 2 e quindi possiamo scegliere una qualsiasi base ortonormale e^1, e^2 di $\mathbb{R}_2$ come assi principali e quindi anche le componenti principali non sono univocamente determinate.

La formula di proiezione

Nota 18.1. Per il lemma 17.9 ogni retta di regressione ortogonale di X passa per il baricentro $\overline{X}$ dei punti X^i e può quindi essere scritta nella forma

$$R_g := \overline{X} + \mathbb{R}g$$

con $g \in \mathbb{R}_m$ e $|g| = 1$. Per la proposizione 4.1 la proiezione ortogonale P_g^i di X^i su R_g è uguale a

$$P_g^i = \overline{X} + \|X^i - \overline{X}, g\|g$$

Come nel lemma 17.9 dal teorema di Pitagora abbiamo

$$|X^i - P_g^i|^2 = |X^i - \overline{X}|^2 - |P_g^i - \overline{X}|^2$$

Ma $|P_g^i - \overline{X}|^2 = \|X^i - \overline{X}, g\|^2$, cosicché

$$\sum_{i=1}^n |X^i - P_g^i|^2 = \sum_{i=1}^n |X^i - \overline{X}|^2 - \sum_{i=1}^n \|X^i - \overline{X}, g\|^2$$

La somma $\sum_{i=1}^n |X^i - \overline{X}|^2$ però dipende solo dalla matrice dei dati X e non dal vettore g che vogliamo variare per ottenere il minimo di $\sum_{i=1}^n |X^i - P_g^i|^2$ e vediamo che quest'ultima somma è minima se e solo se $\sum_{i=1}^n \|X^i - \overline{X}, g\|^2$ è massima.

La matrice di covarianza

Definizione 18.2. Poniamo $\Omega := \Omega_X := (CX)^t CX$.

In statistica la matrice $\mathrm{cov}(X) := \dfrac{\Omega}{n-1}$ si chiama la *matrice di covarianza* di X. Entrambe le matrici appartengono ad $\mathbb{R}_m^m$ e sono simmetriche.

Possiamo definire un operatore $\varphi_X := \bigcirc_f f\Omega : \mathbb{R}_m \longrightarrow \mathbb{R}_m$, evidentemente simmetrico. A questo operatore si riferisce nel seguito il rapporto di Rayleigh:

$$\mathcal{R}g = \|g\Omega, g\|$$

per $g \in \mathbb{R}_m$ con $|g| = 1$.

Lemma 18.3. $\Omega = X^t CX = X^t X - X^t MX$.

Dimostrazione. Chiaramente $C^t = C$, mentre sappiamo dal corollario 2.10 che $C^2 = C$. Perciò

$$\begin{aligned}\Omega = (CX)^t CX = X^t C^t CX = X^t CX \\ = X^t(\delta - M)X = X^t X - X^t MX\end{aligned}$$

Osservazione 18.4.

$$\Omega_j^i = \|CX_i, CX_j\|$$
$$\mathrm{cov}(X)_j^i = s_{X_i X_j}$$

Dimostrazione. Per il corollario 14.10 abbiamo

$$\Omega_j^i = ((CX)^t CX)_j^i = \|(CX)_i, (CX)_j\| = \|CX_i, CX_j\|$$

Dividendo per $n-1$ otteniamo la seconda equazione.

Osservazione 18.5. $(CX)^i = X^i - \overline{X}$ *per ogni* i.

Nota 18.6. Sia $g \in \mathbb{R}_m$ con $|g| = 1$. Per la proposizione 14.14 abbiamo

$$\sum_{i=1}^n \|X^i - \overline{X}, g\|^2 = \|g\Omega, g\| = \mathcal{R}g$$

Come nella nota 18.1 sia P_g^i la proiezione ortogonale di X^i sulla retta $\overline{X} + \mathbb{R}g$. Da quella stessa nota segue allora

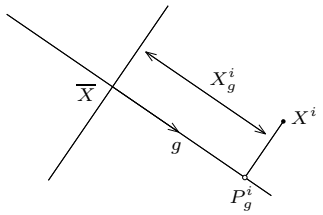
$$\begin{aligned}\min\{\sum_{i=1}^n |X^i - P_g^i|^2 \mid g \in \mathbb{R}_m \text{ con } |g| = 1\} \\ = \max\{\mathcal{R}g \mid g \in \mathbb{R}_m \text{ con } |g| = 1\} \\ = \max \mathcal{R}\mathbb{R}_m\end{aligned}$$

Nota 18.7. $\lambda_1, \ldots, \lambda_m$ siano gli autovalori (necessariamente reali) di Ω e $\lambda_1 \geq \ldots \geq \lambda_m$. Per la nota 12.16 esiste una base ortonormale $e^1, \ldots, e^m$ di $\mathbb{R}_m$ tale che

$$e^1\Omega = \lambda_1 e^1, \ldots, e^m\Omega = \lambda_m e^m$$

Per la nota 18.6 e il corollario 13.8 $\overline{X} + \mathbb{R}e^1$ è una retta di regressione ortogonale di X.

Definizione 18.8. Per $g \in \mathbb{R}_m$ sia $X_g^i := \|X^i - \overline{X}, g\|$ per ogni $i = 1, \ldots, n$ la lunghezza segnata della proiezione ortogonale di $X^i - \overline{X} = (CX)^i$ sulla retta orientata generata da g. Otteniamo in questo modo un vettore colonna $X_g \in \mathbb{R}^n$.



Osservazione 18.9. *Sia* $g \in \mathbb{R}_m$. *Allora* $X_g = CXg^t$.

Dimostrazione. Per ogni i abbiamo
$(CXg^t)^i = (CX)^i g^t = \|(CX)^i, g\| = X_g^i$

Corollario 18.10. *Sia* $g \in \mathbb{R}_m$. *Allora* $\overline{X_g} = 0$.

Dimostrazione. Applicando l'osservazione 2.4 ad X_g abbiamo $MX_g = \overline{X_g}^{\diamond}$. È quindi sufficiente dimostrare che $MX_g = 0$.

Ma $MX_g = MCXg^t = 0$ per l'osservazione 2.11.

Una dimostrazione diretta è altrettanto facile: Abbiamo

$$n\overline{X_g} = \sum_{i=1}^n \|X^i - \overline{X}, g\| = \|\sum_{i=1}^n (X^i - \overline{X}), g\|$$

Ma $\sum_{i=1}^n (X^i - \overline{X}) = 0$.

Corollario 18.11. *Sia* $g \in \mathbb{R}_m$ *con* $|g| = 1$. *Allora*

$$|X_g|^2 = \mathcal{R}g$$
$$s_{X_g}^2 = \frac{1}{n-1}\mathcal{R}g$$

Dimostrazione. $|X_g|^2 = \sum_{i=1}^n \|X^i - \overline{X}, g\|^2 = \mathcal{R}g$

come sappiamo dalla nota 18.6.
La seconda equazione segue adesso dal corollario 18.10.

Nota 18.12. Nelle ipotesi della nota 18.7 poniamo

$$\mathcal{G}_1 := \{g \in \mathbb{R}_m \mid |g| = 1\}$$

e per $2 \leq k \leq m$

$$\mathcal{G}_k := \mathcal{G}_1 \cap E_{1,k-1}^{\perp}$$

Dai corollari 18.12 e 13.8 con la nota 18.6 segue che

$$s_{X_{e^1}}^2 = \max\{s_{X_g}^2 \mid g \in \mathbb{R}_m \text{ e } |g| = 1\} = \frac{\lambda_1}{n-1}$$

In questo senso con e^1 abbiamo scoperto una direzione di massima varianza (non univocamente determinata).

Dalla nota 18.1 e dal corollario 14.3 vediamo che in generale, per $1 \leq k \leq m$, $\overline{X} + \mathbb{R}e^k$ è una retta che minimizza $\sum_{i=1}^n |X^i - P_g^i|^2$ per $g \in \mathcal{G}_k$ e che $s_{X_{e^k}}^2 = \max\{s_{X_g}^2 \mid g \in \mathcal{G}_k\} = \dfrac{\lambda_k}{n-1}$.

I vettori $X_{e^1}, \ldots, X_{e^m}$ sono detti *componenti principali* di X; non sono comunque univocamente determinati, come abbiamo visto nella nota 17.10.

T. Lehmann/W. Oberschelp/E. Pelikan/R. Repges: Bildverarbeitung für die Medizin. Springer 1997.

T. Anderson: An introduction to multivariate statistical analysis. Wiley 2003.

Componenti principali

Situazione 19.1. $X \in \mathbb{R}^n_m$ sia la nostra matrice di dati con $n \geq 2$. Come nella nota 18.7 e quando non indicato diversamente, siano $\lambda_1, \ldots, \lambda_m$ gli autovalori di Ω con $\lambda_1 \geq \ldots \geq \lambda_m$ ed $e^1, \ldots, e^m$ una base ortonormale di $\mathbb{R}_m$ tale che

$$e^1\Omega = \lambda_1 e^1, \ldots, e^m\Omega = \lambda_m e^m$$

Inoltre sia $1 \leq q \leq m$.

Nota 19.2. Sia $g \in \mathbb{R}_m$ con $|g| = 1$. Per definizione

$$X^i_g = \|X^i - \overline{X}, g\|$$

è la lunghezza con segno (calcolata a partire dal baricentro $\overline{X}$) della proiezione ortogonale del punto X^i sulla retta generata da g, come già osservato nella nota 18.7.

Per l'osservazione 18.9 abbiamo $X_g = CXg^t$.

Questa, per la nota 13.16, è una combinazione lineare delle colonne di CX con i coefficienti g_j:

$$X_g = \sum_{j=1}^{m} (CX)_j g_j$$

che, tenendo conto della definizione di CX, possiamo scrivere in forma ancora più esplicita:

$$X_g = \sum_{j=1}^{m} (X_j - \overline{X_j}^{\diamond}) g_j$$

In particolare abbiamo

$$X_{e^k} = \sum_{j=1}^{m} (CX)_j e^k_j = \sum_{j=1}^{m} (X_j - \overline{X_j}^{\diamond}) e^k_j$$

per ogni $k = 1, \ldots, m$.

La k-esima componente principale di X è quindi una combinazione lineare delle colonne di CX con i coefficienti dell'autovettore e^k.

Osservazione 19.3. *$X_{e^j} \perp X_{e^k}$ per ogni $j \neq k$.*

Dimostrazione. Siccome per $a, b \in \mathbb{R}^n$ abbiamo $\|a, b\| = \|a^t, b^t\|$, dove il secondo prodotto scalare è calcolato in $\mathbb{R}_n$, possiamo scrivere

$$\begin{aligned}\|X_{e^j}, X_{e^k}\| &= \|CX(e^j)^t, CX(e^k)^t\| = \|e^j(CX)^t, e^k(CX)^t\| \\ &= \|e^j\Omega, e^k\| = \|\lambda_j e^j, e^k\| = \lambda_j \|e^j, e^k\| = 0\end{aligned}$$

Nota 19.4. Consideriamo la matrice di dati

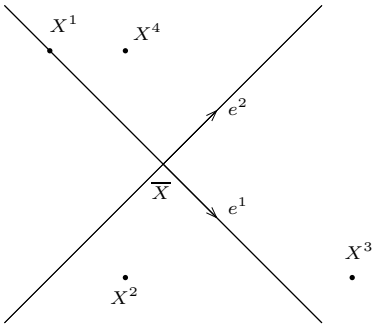
$$X := \begin{pmatrix} 1 & 4 \\ 2 & 1 \\ 5 & 1 \\ 2 & 4 \end{pmatrix}$$

Calcoliamo $\Omega = \begin{pmatrix} 9 & -6 \\ -6 & 9 \end{pmatrix}$.

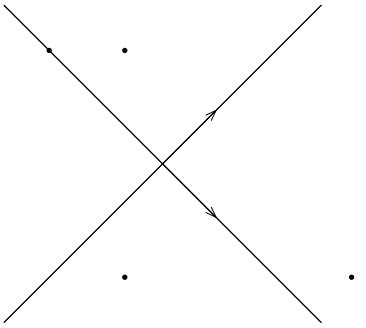
Gli autovalori di Ω sono $\lambda_1 = 15$ e $\lambda_2 = 3$. Ad essi corrispondono gli autovettori ortonormali

$$e^1 = (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}) \quad e^2 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$$

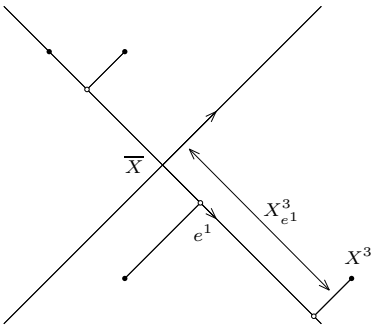
Il baricentro è $\overline{X} = (2.5, 2.5)$; lo troviamo in R con `colMeans(X)` oppure, in un esempio così semplice, con un calcolo diretto.



Vediamo che effettivamente si ha l'impressione che la variazione maggiore avvenga nella direzione e^1, quella minore nella direzione e^2, in accordo con quanto osservato nella nota 18.12. Ciò si vede ancora meglio se togliamo le leggende:



Come osservato, $X^i_{e^1}$ è la lunghezza segnata (calcolata rispetto al centro $\overline{X}$) della proiezione ortogonale di X^i sulla retta generata da e^1:



Nota 19.5. Una volta determinato (in una delle possibili scelte) e^1, otteniamo una proiezione

$$\mathbb{R}_m \longrightarrow \mathbb{R}_1 \text{ con } x \longmapsto \overline{X} + \|x - \overline{X}, e^1\| e^1$$

nella quale in particolare i punti X^i vengono proiettati secondo

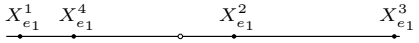
$$X^i \mapsto P^i_{e^1} = \overline{X} + X^i_{e^1} e^1$$

Nell'esempio della nota 19.4 calcoliamo prima le lunghezze delle proiezioni sugli assi principali, ottenendo la matrice

```
-2.12  0.00
 0.71 -1.41
 2.83  0.71
-1.41  0.71
```

in cui la prima colonna si riferisce all'asse determinato da e^1. Usare un righello per misurare (in cm) le lunghezze sull'ultima figura della nota 19.4 per convincersi che il risultato è corretto.

Riportando i valori $X^i_{e^1}$ su un'ascisse otteniamo un'immagine undimensionale dei nostri dati:



Il cerchietto bianco è qui l'origine di $\mathbb{R}_1$.

Sappiamo dalla nota 18.12 che la varianza di questi punti in $\mathbb{R}_1$ è uguale a $\lambda_1 = 15$, mentre la varianza della proiezione sul secondo asse principale è uguale a $\lambda_2 = 3$ e quindi molto minore. Possiamo perciò sperare che X_{e^1} da solo ci dia sufficienti informazioni. Assumiamo che, in un altro esempio, le proiezioni sull'asse più importante siano distribuite come nella seguente figura:

Allora possiamo considerare i punti come appartenenti a tre gruppi distinti e se gli altri autovalori di Ω sono molto più piccoli di λ_1, questo raggruppamento potrà, con molta prudenza, essere considerato significativo. Si tenga conto del fatto che proprio l'analisi delle componenti principali è molto sensibile alla scala usata, ad esempio al cambio delle unità di misura nel rilevamento delle variabili.

Il rapporto di varianza

Ricordiamo dalla nota 18.12 che la varianza di $X_{\cdot k}$ è uguale a λ_k per ogni k e quindi la somma $\lambda_1 + \ldots + \lambda_m$ (nel caso generale) può essere considerata la *varianza totale* dei nostri dati; siccome la *traccia* di una matrice quadratica è uguale alla somma dei suoi autovalori, la varianza totale è uguale alla traccia di Ω. A questo punto è naturale, in una proiezione 2-dimensionale sui primi due assi principali, considerare il quoziente

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_m}$$

detto secondo *rapporto (cumulativo) di varianza*, come indice della bontà statistica della proiezione, da interpretare con molta precauzione, come vedremo, soprattutto quando si confrontano standardizzazioni diverse. Nell'esempio presentato a pagina 30 il rapporto di varianza è uguale a 0.936 e quindi le prime due componenti principali rappresentano più del 93% della varianza totale.

La differenza

$$1 - \frac{\lambda_1 + \lambda_2}{\lambda_1 + \ldots + \lambda_m}$$

è una misura invece della *profondità* dei dati; a una profondità maggiore corrisponde un rischio maggiore che punti vicini nella proiezione sul piano siano invece lontane nella realtà, cioè in $\mathbb{R}_m$.

„The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set ... Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix. Thus, the definition and computation of principal components are straightforward but, as will be seen, this apparently simple technique has a wide variety of different applications, as well as a number of different derivations ... Despite the apparent simplicity of the technique, much research is still being done in the general area of PCA, and it is very widely used.“
(Jolliffe, ix, 9)

Un metodo con molti nomi

L'analisi delle componenti principali appare in molti campi della matematica applicata con diversi nomi: *Trasformazione sugli assi principali* in geometria, *trasformazione di Karhunen-Loève* in ingegneria e nella teoria del riconoscimento delle forme e nell'elaborazione delle immagini, *analisi spettrale* in fisica e analisi matematica (ad esempio problemi agli autovalori per equazioni differenziali), *analisi fattoriale* in psicologia (anche se con questo termine spesso si associano obiettivi più ambiziosi della sola riduzione delle dimensioni). Essa è spesso un primo passo preparatorio che permette di applicare altri metodi della statistica multivariata, come l'analisi dei raggruppamenti e la ricerca di funzioni discriminanti.

In *meccanica* ($m = 3$) la ricerca del primo asse principale (asse con momento inerziale massimo) è importante, perché la rotazione attorno a questo asse gode di stabilità.

Trasformazione affine dei dati

Nota 20.1. Esaminiamo brevemente il comportamento della matrice dei dati quando sottoponiamo le variabili a una trasformazione affine $\bigcirc_u uA + b : \mathbb{R}_m \longrightarrow \mathbb{R}_p$ con $A \in \mathbb{R}_p^m$ e $b \in \mathbb{R}_p$; la matrice opera da destra, perché consideriamo vettori riga.

Da ogni riga X^i della nostra matrice dei dati X otteniamo allora una riga $Y^i = X^i A + b \in \mathbb{R}_p$; possiamo così formare la matrice $Y \in \mathbb{R}_p^n$ formata da queste righe. Si vede facilmente che

$$Y = XA + 1^\diamond b$$

Proposizione 20.2. *Come nella nota 20.1 siano $A \in \mathbb{R}_p^m$ e $b \in \mathbb{R}_p$ ed $Y = XA + 1^\diamond b$. Allora*

$$MY = MXA + 1^\diamond b$$
$$CY = CXA$$

Dimostrazione. La prima equazione segue dalla relazione

$$MY = M(XA + M1^\diamond b) = MXA + M1^\diamond b$$

perché dal corollario 2.5 sappiamo che $M1^\diamond = 1^\diamond$.

Per la seconda equazione abbiamo, usando la prima,

$$\begin{aligned} CY = Y - MY &= XA + 1^\diamond b - (MXA + 1^\diamond b) \\ &= XA - MXA = (X - MX)A = CXA \end{aligned}$$

Varietà di Stiefel

Lemma 20.3. *Sia $G \in \mathbb{R}_m^q$. Allora*

$$\sum_{i=1}^{n} \sum_{j=1}^{q} \|X^i - \overline{X}, G^j\|^2 = \operatorname{tr} G\Omega G^t$$

Dimostrazione. Dal corollario 15.7 abbiamo, usando anche la proposizione 13.14,

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{q} \|X^i - \overline{X}, G^j\|^2 &= \sum_{i=1}^{n} \operatorname{tr} G(X^i - \overline{X})^t (X^i - \overline{X}) G^t \\ &= \operatorname{tr} G(\sum_{i=1}^{n} (X^i - \overline{X})^t (X^i - \overline{X})) G^t = \operatorname{tr} G\Omega G^t \end{aligned}$$

Definizione 20.4. $O(m)$ sia l'insieme delle matrici ortogonali di rango m. È chiaro che una matrice $G \in \mathbb{R}_m^m$ appartiene ad $O(m)$ se e solo se le righe $G^1, \ldots, G^m$ costituiscono una base ortonormale di $\mathbb{R}_m$.

Osservazione 20.5. *Sia $G \in O(m)$. Allora*

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \|X^i - \overline{X}, G^j\|^2 = \operatorname{tr} \Omega$$

Vediamo così che questa espressione non dipende da G.

Dimostrazione. Infatti in questo caso

$$\operatorname{tr} G\Omega G^t = \operatorname{tr} G\Omega G^{-1} = \operatorname{tr} \Omega$$

e l'enunciato segue dal lemma 20.3.

Definizione 20.6. Sia $G \in \mathbb{R}_m^q$. Diciamo che *G possiede righe ortonormali* e scriviamo $G \in O^q(m)$, se le righe di G hanno tutte lunghezza 1 e sono ortogonali tra di loro. Gli insiemi $O^q(m)$ sono noti come *varietà di Stiefel*.

Nota 20.7. Le varietà di Stiefel sono molto importanti in topologia e geometria differenziale. Appaiono anche in modo naturale negli aspetti geometrici della statistica multivariata, non solo nell'analisi delle componenti principali, ma anche ad esempio in algoritmi di proiezione ottimale.

Infatti se vogliamo proiettare punti da $\mathbb{R}_m$ in un piano e variare questo piano, possiamo considerare una curva nella varietà di Grassmann $G^2(m)$, che è l'insieme di tutti i piani (passanti per l'origine) in $\mathbb{R}_m$. Per ottenere le proiezioni abbiamo però bisogno di basi ortonormali in ogni piano e quindi di curve in $O^2(m)$.

Osservazione 20.8. *Sia $G \in \mathbb{R}_m^q$. Allora*

$$G \in O^q(m) \iff GG^t = \delta$$

Si noti che $GG^t \in \mathbb{R}_q^q$ e quindi δ è qui la matrice identica in $\mathbb{R}_q^q$.

Definizione 20.9. $\operatorname{diag}(\lambda_1, \ldots, \lambda_q)$ sia la matrice diagonale i cui coefficienti diagonali sono $\lambda_1, \ldots, \lambda_q$.

Lemma 20.10. *Siano $L \in \mathbb{R}_m^m$, $F \in O^q(m)$ e $\lambda_1, \ldots, \lambda_q \in \mathbb{R}$ tali che $F^i L = \lambda_i F^i$ per ogni i. Allora*

$$FLF^t = \operatorname{diag}(\lambda_1, \ldots, \lambda_q) \text{ e quindi } \operatorname{tr} FLF^t = \sum_{j=1}^{q} \lambda_j$$

Dimostrazione. Per ogni i, j abbiamo

$$(FLF^t)^i_j = (FL)^i (F^t)_j = \lambda_i F^i (F^j)^t = \lambda_i \delta_{ij}$$

usando l'osservazione 13.17.

Lemma 20.11. *Sia $G \in O^q(m)$. Allora* $\operatorname{tr} G^t G = q$.

Dimostrazione. $GG^t = \delta \in \mathbb{R}_q^q$, per cui

$$\operatorname{tr} G^t G = \operatorname{tr} GG^t = \operatorname{tr} \delta = q$$

Ortoregressione su iperpiani

Lemma 21.1. *Sia $G \in O^q(m)$. Allora*

$$\sum_{i=1}^{n}\sum_{j=1}^{q} \|X^i - \overline{X}, g^j\|^2 = \operatorname{tr} G\Omega G^t = \sum_{j=1}^{q} \mathcal{R}G^j$$

Dimostrazione. Per il lemma 20.3 dobbiamo solo dimostrare la seconda uguaglianza. Siccome $|G^j| = 1$ per ogni j, abbiamo

$$\mathcal{R}G^j = \|G^j\Omega, G^j\| = G^j(G^j\Omega)^t = G^j\Omega(G^t)_j = (G\Omega G^t)^j_j$$

per cui $\sum_{j=1}^{q} \mathcal{R}G^j = \operatorname{tr} G\Omega G^t$.

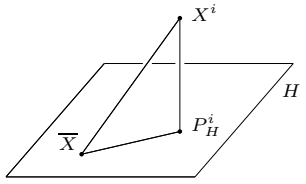
Nota 21.2. Vogliamo adesso dimostrare che, per $1 \le q \le m$, l'iperpiano $\overline{X} + \mathbb{R}e^1 + \ldots + \mathbb{R}e^q$ minimizza la somma dei quadrati delle distanze dei punti X^i da un iperpiano q-dimensionale H di $\mathbb{R}_m$. Chiamiamo un iperpiano con questa proprietà un *q-iperpiano di regressione ortogonale* per X; si può dimostrare anche in questo caso che esso passa per il baricentro $\overline{X}$. Possiamo quindi trovare vettori $g^1, \ldots, g^q \in \mathbb{R}_m$ con $\|g^j, g^k\| = \delta_{jk}$ per ogni j, k e tali che

$$H = \overline{X} + \mathbb{R}g^1 + \ldots + \mathbb{R}g^q$$

Per la proposizione 4.1 la proiezione P^i_H di X^i su H è data da

$$P^i_H = \overline{X} + \sum_{j=1}^{q} \|X^i - \overline{X}, g^j\| g^j$$

Noi dobbiamo scegliere $g^1, \ldots, g^q$ in modo da minimizzare la somma $\sum_{i=1}^{n} |X^i - P^i_H|^2$.



Anche qui abbiamo $|X^i - P^i_H|^2 = |X^i - \overline{X}|^2 - |P^i_H - \overline{X}|^2$.

Ma $|P^i_H - \overline{X}|^2 = \sum_{j=1}^{q} \|X^i - \overline{X}, g^j\|^2$, cosicché

$$\sum_{i=1}^{n} |X^i - P^i_H|^2 = \sum_{i=1}^{n} |X^i - \overline{X}|^2 - \sum_{i=1}^{n}\sum_{j=1}^{q} \|X^i - \overline{X}, g^j\|^2$$

La somma $\sum_{i=1}^{n} |X^i - \overline{X}|^2$ però dipende solo dalla matrice dei dati X e non dai vettori g^j che vogliamo variare per ottenere il minimo di $\sum_{i=1}^{n} |X^i - P^i_H|^2$ e vediamo che quest'ultima somma è minima se e solo se $\sum_{i=1}^{n}\sum_{j=1}^{q} \|X^i - \overline{X}, g^j\|^2$ è massima.

Dal lemma 21.1 vediamo che dobbiamo trovare

$$\max\{\operatorname{tr} G\Omega G^t \mid G \in O^q(m)\} = \max\{\sum_{j=1}^{q} \mathcal{R}G^j \mid G \in O^q(m)\}$$

Osservazione 21.3. *$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ sia una matrice diagonale e $B \in \mathbb{R}^m_m$. Allora* $\operatorname{tr} DB = \sum_{j=1}^{m} \lambda_j B^j_j$.

Dimostrazione. Infatti

$$(DB)^i_j = \sum_{k=1}^{m} D^i_k B^k_j = \sum_{k=1}^{m} \delta_{ik}\lambda_i B^k_j = \lambda_i B^i_j$$

Osservazione 21.4. *$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$ sia una matrice diagonale ed $A \in \mathbb{R}^q_m$. Allora* $\operatorname{tr} ADA^t = \sum_{j=1}^{m} \lambda_j |A_j|^2$.

Dimostrazione. Abbiamo

$$\operatorname{tr} ADA^t = \operatorname{tr} DA^tA \overset{21.3}{=} \sum_{j=1}^{m} \lambda_j (A^tA)^j_j. \text{ Ma } (A^tA)^j_j = |A_j|^2.$$

Osservazione 21.5. *Siano $u \in \mathbb{R}^q$, $v \in \mathbb{R}^s$ e $w := \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^{q+s}$.*

Allora $|w|^2 = |u|^2 + |v|^2$.

Ciò implica in particolare $|u|^2 \le |w|^2$.

Lemma 21.6. *Siano dati numeri reali $x_1, \ldots, x_m$, $\lambda_1, \ldots, \lambda_m$ tali che siano soddisfatte le seguenti condizioni:*

(1) $0 \le x_j \le 1$ *per ogni j.*

(2) $\sum_{j=1}^{m} x_j = q$.

(3) $\lambda_1 \ge \ldots \ge \lambda_m$.

Allora $\sum_{j=1}^{q} \lambda_j \ge \sum_{j=1}^{m} \lambda_j x_j$.

Dimostrazione. L'enunciato può essere interpretato come l'affermazione che il compito di ottimizzazione

$$Q(x_1, \ldots, x_m) = \max \text{ con } Q(x_1, \ldots, x_m) := \sum_{j=1}^{m} \lambda_j x_j$$

sotto le condizioni (1) e (2) possiede la soluzione

$$x_1 = \ldots = x_q = 1, x_{q+1} = \ldots = x_m = 0$$

Ciò è però evidente, perché significa che dobbiamo concentrare le „risorse“ x_j nei primi q posti dove la rendita è massima, esaurendo in questo modo però la risorsa totale q.

Osservazione 21.7. Con $E = (e^1, \ldots, e^m)^t$ si ha

$$E\Omega E^t = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$$

Ciò, come nel lemma 20.10, è immediato dalle relazioni $e^i\Omega = \lambda_i e^i$ per ogni i.

Teorema 21.8. *Ponendo ancora $E := \begin{pmatrix} e^1 \\ \vdots \\ e^m \end{pmatrix}$ sia $F \in \mathbb{R}^q_m$ la matrice che consiste delle prime q righe di E.*

Allora $F \in O^q(m)$ e per ogni $G \in O^q(m)$ vale

$$\operatorname{tr} F\Omega F^t \ge \operatorname{tr} G\Omega G^t$$

Dimostrazione. Sia $G \in O^q(m)$.

Siccome le righe di E costituiscono una base di $\mathbb{R}_m$, per ogni i esiste una rappresentazione $G^i = \sum_{j=1}^{m} H^i_j E^j$. I coefficienti H^i_j formano una matrice $H \in \mathbb{R}^q_m$ per cui $G = HE$.

Sia $D := \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$.

Per l'osservazione 21.7 abbiamo $G\Omega G^t = HE\Omega E^t H^t = HDH^t$,

cosicché dall'osservazione 21.4 segue $\operatorname{tr} G\Omega G^t = \sum_{j=1}^{m} \lambda_j |H_j|^2$.

Abbiamo inoltre $H = GE^{-1} = GE^t$ e quindi

$$HH^t = GE^tEG^t = GG^t = \delta$$

per cui $H \in O^q(m)$. Esiste perciò una matrice $K \in O^{m-q}(m)$ tale che $U := \begin{pmatrix} H \\ K \end{pmatrix} \in O(m)$.

Per l'osservazione 21.5 ciò implica $|H_j|^2 \le |U_j|^2 = 1$ per ogni j.

Per il lemma 20.11 $\operatorname{tr} H^tH = q$ e quindi, per la nota 15.8, $\sum_{j=1}^{m} |H_j|^2 = q$. Per il lemma 20.10 $\operatorname{tr} F\Omega F^t = \sum_{j=1}^{q} \lambda_j$.

L'enunciato segue dal lemma 21.6.

Teorema 21.9. *$\overline{X} + \mathbb{R}e^1 + \ldots + \mathbb{R}e^q$ è un q-iperpiano di regressione ortogonale per X.*

Dimostrazione. Teorema 21.8 e nota 21.2.

B. Flury: A first course in multivariate statistics. Springer 1997.
J. Gentle: Elements of computational statistics. Springer 2002.
I. Jolliffe: Principal component analysis. Springer 2002.
K. Mardia/J. Kent/J. Bibby: Multivariate analysis. Academic Press 2000.

V. PROGRAMMAZIONE IN R

R ed S-Plus

R è un linguaggio di programmazione ad altissimo livello orientato soprattutto all'uso in statistica. In verità lo sbilanciamento verso la statistica non deriva dalla natura del linguaggio, ma dalla disponibilità di grandi raccolte di funzioni statistiche e dagli interessi dei ricercatori che lo hanno inventato e lo mantengono. R è gratuito e molto simile a un linguaggio commerciale, S, creato negli anni '80 e anch'esso molto usato. S viene commercializzato come sistema S-Plus. Le differenze non sono grandissime se non sul piano della programmazione, dove R aderisce a una impostazione probabilmente più maneggevole.

R ed S-Plus sono particolarmente popolari nella statistica medica, ma vengono anche usati nella statistica economica o sociale, in geografia, nella matematica finanziaria. L'alto livello del linguaggio permette di creare facilmente librerie di funzioni per nuove applicazioni. Il punto debole è la velocità di esecuzione in calcoli numerici in grandi dimensioni, mentre sono ricchissime le capacità grafiche.

Benché così indirizzato verso la statistica, R non deve essere considerato un pacchetto di statistica. È un vero linguaggio di programmazione, anzi un linguaggio di programmazione molto avanzato, e ciò permette di adattarlo ad ogni compito informatico. Nella stessa statistica questa flessibilità è molto importante proprio oggi, dove continuamente si scoprono nuovi bisogni applicativi, nuove necessità di tradurre metodi matematici, ad esempio nella statistica di complessi dati clinici o geografici, in strumenti informatici.

Un'introduzione alla programmazione in R si trova nel corso di Fondamenti di informatica 2004/05 e, per quanto riguarda la grafica, nel corso di Algoritmi e strutture di dati 2004/05.

Python

Python è in questo momento forse il miglior linguaggio di programmazione: per la facilità di apprendimento e di utilizzo, per le caratteristiche di linguaggio ad altissimo livello che realizza i concetti sia della programmazione funzionale che della programmazione orientata agli oggetti, per il recente perfezionamento della libreria per la programmazione insiemistica, per il supporto da parte di numerosi programmatori, per l'ottima documentazione disponibile in rete e la ricerca riuscita di meccanismi di leggibilità, per la grafica con Tkinter, per la semplicità dell'aggancio ad altri linguaggi, di cui il modulo RPy per il collegamento con R è un esempio meraviglioso.

Enthought Python è una raccolta molto ricca (di 124 MB compressi) per Windows che non comprende soltanto il linguaggio Python (attualmente nella versione 2.4.3), ma anche numerose librerie scientifiche e grafiche, particolarmente utili per il matematico.

Un'introduzione alla programmazione in Python si trova nel corso di Programmazione 05/06.

Esecuzione di un programma in Python

In una sottocartella apposita creiamo i files sorgente come files di testo puro con l'estensione *.py*, utilizzando l'editor incorporato di Python. Con lo stesso editor, piuttosto comodo, scriviamo anche, usando l'estensione *.r*, le sorgenti in R che vogliamo affiancare ai programmi in Python. I programmi possono essere eseguiti mediante il tasto *F5* nella finestra dell'editor. Soprattutto in fase di sviluppo sceglieremo questa modalità di esecuzione, perché così vengono visualizzati anche i messaggi d'errore.

Successivamente i programmi possono essere eseguiti anche tramite il clic sull'icona del file oppure, in un terminale (prompt dei comandi) e se il file si chiama *alfa.py*, con il comando `python alfa.py`.

Per importare le istruzioni contenute in un file `beta.py` si usa il comando `import beta`, tralasciando l'estensione. Con lo stesso comando si importano i moduli delle librerie incorporate o prelevate in rete:

```
import os, math, scipy
```

Teoricamente i programmi possono essere scritti con un qualsiasi editor che crea files in formato testo puro, ad esempio il *Blocco note* di Windows, ma preferiamo utilizzare l'editor di Python per una più agevole correzione degli errori, per l'indentazione automatica e perché prevede la possibilità di usare combinazioni di tasti più comode di quelle disponibili per il Blocco note.

Utilizzo di RPy

Il modulo RPy è un piccolo miracolo e permette una quasi perfetta e semplicissima collaborazione tra Python ed R.

Sotto Linux il pacchetto va installato nel modo seguente: In primo luogo è necessario che R sia stato creato in modo che si possano utilizzare le librerie condivise:

```
./configure --enable-R-shlib
make
make install
```

Successivamente va aggiunta la riga

```
/usr/local/lib/R/lib
```

nel file */etc/ld.so.conf* ed eseguito il comando `ldconfig`.

A questo punto, per installare RPy stesso, è sufficiente

```
/usr/local/bin/python setup.py install
```

Per importare il pacchetto scriviamo

```
from rpy import r
```

nel programma in Python. Le funzioni di R possono allora essere usate anteponendo il prefisso `r.`, come in `r.fun(argomenti)`.

Sotto Linux bisogna (per un piccolo errore contenuto nel modulo) inserire l'istruzione `r.q()` alla fine del programma.

Definiamo ad esempio una funzione in Python che utilizza la funzione `mean` di R per calcolare la media di un vettore:

```
def Media (x): return r.mean(x)
```

Per provare la funzione usiamo

```
print Media([1,5,8,6,3,1])
# output: 4.0
```

In particolare possiamo usare la funzione `source` di R. Ciò significa che possiamo creare una raccolta di funzioni in R da noi programmate; queste funzioni possono a loro volta utilizzare (come se fossimo in una libreria creata per R) le altre funzioni di quella raccolta e allo stesso tempo essere usate, nella sintassi indicata, nei programmi in Python! Creiamo ad esempio un file *funz.r*:

```
# funz.r

cubo = function (x): x^2
```

In uno script di Python scriviamo poi

```
r.source('funz.r')

print r.cubo(13)
# output: 2197.0
```

Installazione di R e di Python

Dal sito del corso installare nell'ordine indicato:

Enthought Python 2.4.3
R 2.2.1
pywin32-209.win32-py2.4.exe
rpy-0.4.6-R-2.0.0-to-2.2.1-py24.win32.exe

Attualmente per Windows sono queste le versioni compatibili con l'utilizzo di RPy anche se di R esiste già la versione 2.3.

Programmi elementari in Python

```
a=range(5,13,2)
print a
# output: [5, 7, 9, 11]

a=range(5,13)
print a
# output: [5, 6, 7, 8, 9, 10, 11, 12]

a=range(13)
print a
# output: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
```

Si noti che il limite destro non viene raggiunto.

```
a=xrange(5,13,2)
print a
# output: xrange(5, 13, 2)

for x in a: print x,
# output: 5 7 9 11
```

La differenza tra `range` e `xrange` è questa: Mentre `range(1000000)` genera una lista di un milione di elementi, `xrange(1000000)` crea questi elementi uno allo volta in ogni passaggio di un ciclo in cui il comando viene utilizzato.

Si noti il doppio punto (`:`) alla fine del comando `for`.

Sono possibili assegnazioni e confronti simultanei:

```
if 3<5<9: print 'o.k.'
# output: o.k.

a=b=c=4
for x in [a,b,c]: print x,
print
# output: 4 4 4
```

Funzioni in Python:

```
def f(x): return 2*x+1

def g(x):
  if (x>0): return x
  else: return -x

for x in xrange(0,10): print f(x),
print
# output: 1 3 5 7 9 11 13 15 17 19

for x in xrange(-5,5): print g(x),
print
# output: 5 4 3 2 1 0 1 2 3 4
```

La virgola alla fine di un comando `print` fa in modo che la stampa continui sulla stessa riga. Come si vede nella definizione di `g`, Python utilizza l'indentazione per strutturare il programma. Anche le istruzioni `if` ed `else` richiedono il doppio punto.

Una funzione di due variabili:

```
def raggio (x,y): return x**2+y**2

print raggio(2,3)
# output: 13
```

Funzioni possono essere non solo argomenti, ma anche risultati di altre funzioni:

```
def sommax (f,g,x): return f(x)+g(x)

def compx (f,g,x): return f(g(x))

def u(x): return x**2

def v(x): return 4*x+1

print sommax(u,v,5)
# output: 46

print compx(u,v,5)
# output: 441
```

Possiamo però anche definire

```
def somma (f,g):
  def s(x): return f(x)+g(x)
  return s

def comp (f,g):
  def c(x): return f(g(x))
  return c

def u(x): return x**2

def v(x): return 4*x+1

print somma(u,v)(5)
# output: 46

print comp(u,v)(5)
# output: 441
```

Queste costruzioni significano che Python appartiene (come R) alla famiglia dei potenti linguaggi *funzionali*.

Stringhe sono racchiuse tra apici o virgolette, stringhe su più di una riga tra triplici apici o virgolette:

```
print 'Carlo era bravissimo.'
# output: Carlo era bravissimo.

print "Carlo e' bravissimo."
# output: Carlo e' bravissimo.

print '''Stringhe a piu' righe si
usano talvolta nei commenti.'''
# output:
# Stringhe a piu' righe si
# usano talvolta nei commenti.
```

Funzioni con un numero variabile di argomenti: Se una funzione è dichiarata nella forma `def f(x,y,*a):`, l'espressione `f(2,4,5,7,10,8)` viene calcolata in modo che gli ultimi argomenti vengano riuniti in una tupla `(5,7,10,8)` che nel corpo del programma può essere vista come se questa tupla fosse a.

```
def somma (*a):
  s=0
  for x in a: s+=x
  return s

print somma(1,2,3,10,5)
# output: 21
```

Ricordiamo che lo schema di Horner per il calcolo dei valori $f(\alpha)$ di un polinomio $f = a_0x^n + \ldots + a_n$ consiste nella ricorsione

$$b_{-1} = 0$$
$$b_k = \alpha b_{k-1} + a_k$$

per $k = 0, \ldots, n$. Possiamo quindi definire

```
def horner (alfa,*a):
  alfa=float(alfa); b=0
  for t in a: b=b*alfa+t
  return b

print horner(10,6,2,0,8)
# output: 6208.0
```

Vettori associativi (dizionari o tabelle di hash) vengono definiti nel modo seguente:

```
latino = {'casa': 'domus', 'villaggio': 'pagus',
  'nave':'navis', 'campo':'ager'}
voci=sorted(latino.keys())
for x in voci: print '%-9s = %s' %(x,latino[x])
# output:
# campo     = ager
# casa      = domus
# nave      = navis
# villaggio = pagus
```

Scambi: `a=5; b=3; [a,b]=[b,a]; print [a,b] # output: [3, 5]`

Programmi elementari in R

```
a=c(1,2,3,7:10)
print(a)
# output: 1 2 3 7 8 9 10

a=seq(2,10,by=3)
print(a)
# output: 2 5 8

a=seq(0,1.5,length=7)
print(a)
# output: 0.00 0.25 0.50 0.75 1.00 1.25 1.50

a=seq(4,by=0.2,length=6)
print(a)
# output: 4.0 4.2 4.4 4.6 4.8 5.0
```

`a:b` è un'abbreviazione per `seq(a,b)`.

Funzioni in R - la sintassi è piuttosto discutibile:

```
f = function (x) 2*x+1

g = function (x)
{if (x>0) x else -x}

for (x in seq(0,9)) cat(f(x),'-',sep='')
cat('\n')
# output: 1-3-5-7-9-11-13-15-17-19-

for (x in seq(-5,5)) cat(g(x),' ',sep='')
cat('\n')
# output: 5 4 3 2 1 0 1 2 3 4 5
```

L'ultimo valore calcolato è il risultato della funzione; non è necessario il return. Una funzione in tre variabili:

```
f = function (x,y,z) x+y^z

print(f(5,4,3))
# output: 69
```

Anche R è un linguaggio funzionale:

```
sommax = function (f,g,x) f(x)+g(x)

compx = function (f,g,x) f(g(x))

u = function (x) x^2

v = function (x) 4*x+1

print(sommax(u,v,5))
# output: 46

print(compx(u,v,5))
# output: 441

somma = function (f,g)
  function (x) f(x)+g(x)

comp = function (f,g)
  function (x) f(g(x))

y=somma(u,v)(5)
print(y)
# output: 46

y=comp(u,v)(5)
print(y)
# output: 441
```

R è un linguaggio prettamente vettoriale; da questo punto di vista è superiore a Python:

```
mediaf = function (f)
  function (x) mean(f(x))

quad = function (x) x^2

m=mediaf(quad)(1:4)
print(m)
# output: 7.5
```

Funzioni con un numero variabile di argomenti:

```
somma = function (...)
# superflua, perche' esiste sum.
{s=0; a=list(...)
for (x in a) s=s+x; s}

y=somma(1,2,3,10,5)
print(y)
# output: 21

y=sum(1,2,3,10,5)
print(y)
# output: 21
```

Si noti che R *non* usa l'indentazione!

Usando la vettorialità di R, spesso si possono evitare cicli. Ciò rende i programmi molto più veloci.

Per creare una tavola dei quadrati dei numeri da 1 a 100 non useremo

```
tav=c()
for (n in 1:100) tav[n]=n^2
```

ma semplicemente

```
n=1:100; tav=n^2
```

R possiede dei meccanismi molto generali e sofisticati per l'uso degli indici in vettori e matrici. Consideriamo il vettore

```
v=11:18
```

Indicando un singolo indice o un vettore di indici ne possiamo estrarre singoli elementi o parti:

```
w=v[2]
print(w)

w=v[c(2,3,8)]
print(w)

w=v[5:9]
print(w)
```

con output

```
12
12 13 18
15 16 17 18 NA
```

Mediante l'uso di indici negativi possiamo escludere alcuni elementi:

```
w=v[-2]
print(w)

w=v[-c(2,3,8)]
print(w)
```

ottenendo

```
11 13 14 15 16 17 18
11 14 15 16 17
```

Una caratteristica avanzata di R è che come indici si possono anche usare vettori di valori booleani, cioè vettori i cui componenti sono o `T` o `F`. Se in questo caso il vettore booleano ha una lunghezza minore di quella del vettore da cui vogliamo estrarre, i valori booleani vengono ciclicamente ripetuti. Esempi:

```
v=c(1:8)
filtro=c(F,F,T,T,F,T,F,F)
u=v[filtro]
print(u)
```

con output `3 4 6`. Infatti vengono riprodotti in `u` gli elementi di `v` che corrispondono a posizioni in cui il valore del vettore booleano è uguale a `T`.

Con `filtro=c(T,F)` otteniamo ogni secondo elemento di `v`, con `filtro=c(F,T)` ogni secondo elemento di `v`, saltando il primo.

apply in R

Introduciamo qui una funzione di R fondamentale per la trasformazione di righe o colonne di una matrice. Siano `f` una funzione definita per vettori (a valori non necessariamente numerici) che per ogni argomento restituisca un vettore della stessa lunghezza ≥ 1 ed `A` una matrice (a valori non necessariamente numerici). Allora

```
t(apply(A,1,f,***))
```

è la matrice che si ottiene da `A` eseguendo `f` su ogni riga di `A`, ed

```
apply(A,2,f,***)
```

è la matrice che si ottiene eseguendo `f` su ogni colonna di `A`. In entrambi i casi `***` indica eventuali ulteriori argomenti di `f`.

Esempio:

```
A = matrix (c(1:4,2:9,3:7,16:9,1:3),ncol=4)
print(A)

# 1 5  5 12
# 2 6  6 11
# 3 7  7 10
# 4 8 16  9
# 2 9 15  1
# 3 3 14  2
# 4 4 13  3

B = apply(A,2,sort)
print(B)

# 1 3  5  1
# 2 4  6  2
# 2 5  7  3
# 3 6 13  9
# 3 7 14 10
# 4 8 15 11
# 4 9 16 12
```

Definiamo le seguenti due funzioni: `Smg.M(X)` calcola MX, `Smg.cen(X)` corrisponde a CX. Si noti comunque che la seconda è più semplice e, a causa della vettorialità delle operazioni in R, non richiede la prima.

```
Smg.M = function (X)
{n=nrow(X)
apply(X,2,function (x) rep(mean(x),n))}

Smg.cen = function (X)
apply(X,2,function (x) x-mean(x))
```

`rep(a,n)` è un vettore che consiste di n copie di a. Anche questa funzione può essere usata in diverse variazioni, ad esempio

```
rep(1:4,c(1,3,2,7))
# 1 2 2 2 3 3 4 4 4 4 4 4 4
```

apply in Python

In Python `apply` è più semplice e corrisponde infatti ad `lapply` in R. f sia una funzione di n argomenti, dove n può essere anche variabile, e v una sequenza di lunghezza n. Allora `apply(f,v)` è uguale ad $f(v_1, \ldots, v_n)$. Esempi:

```
def somma (*a):
  s=0
  for x in a: s+=x
  return s

v=[1,2,4,9,2,8]
s=apply(somma,v)
print s # 26
```

Commenti

Sia in R che in Python, ma anche in molti altri linguaggi interpretati (Perl, la shell di Unix), se una riga contiene, al di fuori di una stringa, il carattere `#`, tutto il resto della riga è considerato un *commento*, compreso il carattere `#` stesso.

In C e C++ una funzione analoga è svolta dalla sequenza `//`.

Autovalori con R

In R gli autovalori di una matrice reale o complessa si trovano con la funzione `eigen`. Essa calcola, se non si pone l'opzione

```
only.values=T
```

anche un sistema di autovettori; ciò può rallentare il calcolo per matrici molto grandi. Il risultato è una *lista* in R e le componenti si ottengono con la sintassi `$values` e `$vectors`. Creiamo due funzioni per la nostra libreria:

```
Mm.autovalori = function (A,simm=F)
{eigen(A,only.values=T,
  symmetric=simm)$values}

Mm.autovettori = function (A,simm=F)
{eigen(A,symmetric=simm)$vectors}
```

`Mm.autovalori` restituisce la lista degli autovalori di `A`, in ordine decrescente (rispetto al modulo se complessi).

`Mm.autovettori` restituisce una *matrice* le cui colonne sono autovettori di `A` corrispondenti agli autovalori nell'ordine indicato. Con l'opzione `simm=T` possiamo accelerare i calcoli per matrici simmetriche.

Per il calcolo di Ω potremmo usare la definizione 18.2 e la funzione `Smg.cen` definita nella colonna a sinistra; possiamo però anche usare la relazione

$$\Omega = (n-1)\operatorname{cov}(X)$$

con

```
Sm.Omega = function (X)
(nrow(X)-1)*cov(X)
```

Siccome `Mm.autovettori(Ω)` è una matrice le cui colonne sono gli autovettori di Ω già normati e nell'ordine desiderato (cioè corrispondenti agli autovalori elencati in ordine decrescente) e tenendo conto del fatto che questi autovettori in R sono vettori colonna, possiamo trovare gli autovettori e i componenti principali di X con le seguenti funzioni:

```
Smp.autovettori = function (X)
{Omega=Sm.Omega(X)
Mm.autovettori(Omega,simm=T)}

Smp = function (X)
Smg.cen(X)%*%Smp.autovettori(X)
```

Potremmo anche usare la funzione `princomp` di R:

```
Smp.R = function (X)
{p=princomp(X)$scores
dimnames(p)=NULL; p}
```

L'istruzione `dimnames(p)=NULL` ha lo scopo di ridurre gli attributi della matrice a quelli di una matrice pura.

Per gli autovalori di Ω usiamo

```
Smp.autovalori = function (X)
{Omega=Sm.Omega(X)
Mm.autovalori(Omega,simm=T)}
```

VI. RAPPRESENTAZIONI GRAFICHE

Quindici comuni

Lavoreremo negli esempi spesso con la seguente tabella di quindici comuni italiani, di cui abbiamo quattro dati: numero degli abitanti, altezza sul mare, distanza dal mare, superficie del territorio comunale. Per avere numeri di grandezza confrontabili, indichiamo gli abitanti in migliaia, l'altezza in metri, la distanza dal mare in chilometri, la superficie in chilometri quadrati.

comune	ab.	alt.	mare	sup.
Belluno	35	383	75	148
Bologna	380	54	70	141
Bolzano	97	262	140	53
Ferrara	132	9	45	405
Firenze	375	50	75	103
Genova	632	19	2	236
Milano	1302	122	108	182
Padova	210	12	25	93
Parma	170	55	90	261
Pisa	92	4	10	188
Ravenna	140	4	8	660
Torino	901	239	105	131
Trento	106	194	110	158
Venezia	275	1	0	458
Vicenza	110	39	55	81

I nomi naturalmente non fanno parte della matrice dei dati che in questo esempio è uguale a

$$X = \begin{pmatrix} 35 & 383 & 75 & 148 \\ 380 & 54 & 70 & 141 \\ 97 & 262 & 140 & 53 \\ 132 & 9 & 45 & 405 \\ 375 & 50 & 75 & 103 \\ 632 & 19 & 2 & 236 \\ 1302 & 122 & 108 & 182 \\ 210 & 12 & 25 & 93 \\ 170 & 55 & 90 & 261 \\ 92 & 4 & 10 & 188 \\ 140 & 4 & 8 & 660 \\ 901 & 239 & 105 & 131 \\ 106 & 194 & 110 & 158 \\ 275 & 1 & 0 & 458 \\ 110 & 39 & 55 & 81 \end{pmatrix}$$

Arrotondando abbiamo

$\overline{X_1} = 330.5 \quad \overline{X_2} = 96.5 \quad \overline{X_3} = 61.2 \quad \overline{X_4} = 219.9$

Possiamo così calcolare

$$CX = M - MX = \begin{pmatrix} -295.5 & 286.5 & 13.8 & -71.9 \\ 49.5 & -42.5 & 8.8 & -78.9 \\ -233.5 & 165.5 & 78.8 & -166.9 \\ -198.5 & -87.5 & -16.2 & 185.1 \\ 44.5 & -46.5 & 13.8 & -116.9 \\ 301.5 & -77.5 & -59.2 & 16.1 \\ 971.5 & 25.5 & 46.8 & -37.9 \\ -120.5 & -84.5 & -36.2 & -126.9 \\ -160.5 & -41.5 & 28.8 & 41.1 \\ -238.5 & -92.5 & -51.2 & -31.9 \\ -190.5 & -92.5 & -53.2 & 440.1 \\ 570.5 & 142.5 & 43.8 & -88.9 \\ -224.5 & 97.5 & 48.8 & -61.9 \\ -55.5 & -95.5 & -61.2 & 238.1 \\ -220.5 & -57.5 & -6.2 & -138.9 \end{pmatrix}$$

Il comune di Ferrara ha un territorio molto grande, corrispondente a un quadrato di 20 km di lato, praticamente uguale a quello di Vienna (415 km^2) e più del doppio di quello di Milano.

I nostri dati si riferiscono circa al 1989; in una lista sulla Wikipedia con dati del 2001 Ferrara figura al 18^o posto tra i comuni italiani elencati per superficie (in km^2): Roma 1285, Ravenna 652, Cerignola 593, Noto 550, Sassari 545, Monreale 529, Gubbio 525, Foggia 507, Grosseto 474, L'Aquila 466, Perugia 449, Ragusa 442, Altamura 427, Caltanissetta 416, Venezia 412, Andria 407, Viterbo 406, Ferrara 404. Al 36^o posto si trova il più grande comune del Trentino Alto Adige, Sarentino con 302 km^2 e 6650 abitanti.

Lettura dei dati con read.table

Conserviamo i dati sui 15 comuni in un file in formato testo che chiamiamo .../*R*/*Dati*/*Quindici-comuni* con il seguente contenuto:

```
"comune" "ab1000" "altmetri" "dalmarekm" "supkm"
"1" "Belluno" 35 383 75 148
"2" "Bologna" 380 54 70 141
"3" "Bolzano" 97 262 140 53
"4" "Ferrara" 132 9 45 405
"5" "Firenze" 375 50 75 103
"6" "Genova" 632 19 2 236
"7" "Milano" 1302 122 108 182
"8" "Padova" 210 12 25 93
"9" "Parma" 170 55 90 261
"10" "Pisa" 92 4 10 188
"11" "Ravenna" 140 4 8 660
"12" "Torino" 901 239 105 131
"13" "Trento" 106 194 110 158
"14" "Venezia" 275 1 0 458
"15" "Vicenza" 110 39 55 81
```

Per trasformare i dati in una *tabella* di R usiamo il comando

```
tab=read.table('Dati/Quindici-comuni')
```

Con `tab` dal terminale di R la tabella viene visualizzata nel modo seguente:

```
     comune ab1000 altmetri dalmarekm supkm
1   Belluno     35      383        75   148
2   Bologna    380       54        70   141
3   Bolzano     97      262       140    53
4   Ferrara    132        9        45   405
5   Firenze    375       50        75   103
6    Genova    632       19         2   236
7    Milano   1302      122       108   182
8    Padova    210       12        25    93
9     Parma    170       55        90   261
10     Pisa     92        4        10   188
11 Ravenna    140        4         8   660
12  Torino    901      239       105   131
13  Trento    106      194       110   158
14 Venezia    275        1         0   458
15 Vicenza    110       39        55    81
```

Possiamo ottenere una matrice di dati da una tabella che, tranne le colonne indicate in `senza`, deve essere omogenea, con la seguente funzione:

```
Dt.matrice = function (tab,senza=1)
{nomi=colnames(tab)
if (!identical(senza,0))
nomi=nomi[-senza]; a=tab[nomi]
v=unlist(a,use.names=F)
matrix(v,ncol=ncol(a))}
```

Nel nostro caso con

```
X=Dt.matrice(tab)
X
```

viene visualizzata la matrice

```
  35  383   75  148
 380   54   70  141
  97  262  140   53
 132    9   45  405
 375   50   75  103
 632   19    2  236
1302  122  108  182
 210   12   25   93
 170   55   90  261
  92    4   10  188
 140    4    8  660
 901  239  105  131
 106  194  110  158
 275    1    0  458
 110   39   55   81
```

Per vedere solo la seconda e la terza colonna dobbiamo usare `Dt.matrice(tab,senza=c(1,2,4))`. Perche?

Proiezione affine su $[0, 1]$

In statistica conviene spesso trasformare i valori contenuti in un vettore `x` di dati numerici in valori compresi tra 0 e 1. Ciò può essere ottenuto con l'operazione affine

$$\bigcirc_{\xi} \frac{\xi - m}{M - m}$$

applicata agli elementi di `x`, dove m è il minimo in `x`, M il massimo. Denotiamo il vettore così ottenuto con x^{01}. Se, come sempre assumiamo, gli elementi di x non sono tutti uguali, allora $m \neq M$.

In R programmiamo

```
S.tra01 = function (v)
{m=min(v); (v-m)/(max(v)-m)}
```

Questa funzione fa parte della sezione S (statistica) della nostra libreria. Esempio:

```
x=1:5
print(S.tra01(x))
# 0.00 0.25 0.50 0.75 1.00
```

Per applicare `S.tra01` a tutte le colonne di una matrice, ottenendo così la matrice

$$X^{01} := (X_1^{01}, \ldots, X_m^{01})$$

combiniamo questa funzione con `apply`:

```
Sm.tra01 = function (X)
apply(X,2,S.tra01)
```

Per la matrice X dei 15 comuni su questa stessa pagina con `Sm.tra01(X)` otteniamo allora, dopo arrotondamento,

$$\begin{pmatrix}
0.00 & 1.00 & 0.54 & 0.16 \\
0.27 & 0.14 & 0.50 & 0.14 \\
0.05 & 0.68 & 1.00 & 0.00 \\
0.08 & 0.02 & 0.32 & 0.58 \\
0.27 & 0.13 & 0.54 & 0.08 \\
0.47 & 0.05 & 0.01 & 0.30 \\
1.00 & 0.32 & 0.77 & 0.21 \\
0.14 & 0.03 & 0.18 & 0.07 \\
0.11 & 0.14 & 0.64 & 0.34 \\
0.04 & 0.01 & 0.07 & 0.22 \\
0.08 & 0.01 & 0.06 & 1.00 \\
0.68 & 0.62 & 0.75 & 0.13 \\
0.06 & 0.51 & 0.79 & 0.17 \\
0.19 & 0.00 & 0.00 & 0.67 \\
0.06 & 0.10 & 0.39 & 0.05
\end{pmatrix}$$

Questa tecnica è utile molto spesso tranne nei casi in cui, per la presenza di uno o più valori eccezionali in una colonna, la colonna trasformata diventa troppo concentrata su una piccola porzione dell'intervallo $[0, 1]$:

```
x=c(2,3,5,6,7,11,13,100)
x1=S.tra01(x)
print(round(x1,2))
# 0 0.01 0.03 0.04 0.05 0.09 0.11 1
```

Proprio nella statistica esploratoria lo studio di X^{01} è in genere da preferire all'uso della normalizzazione statistica

$$\breve{X} := (\breve{X_1}, \ldots, \breve{X_m})$$

che appartiene piuttosto alla statistica parametrico-inferenziale.

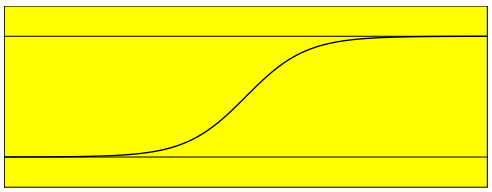
La struttura fondamentale per rappresentare dati statistici in R sono le *tabelle* (in inglese *data frames*). Molte funzioni statistiche di R operano su tabelle. Esse sono simili alle matrici dalle quali si distinguono per il fatto che le colonne possono corrispondere a tipi diversi. Gli elementi in ogni colonna invece sono dello stesso tipo.

Formalmente una tabella è una *lista* di vettori della stessa lunghezza con nomi distinti per le colonne con cui queste possono essere identificate. Più dettagli si trovano nel numero 9 del corso di Fondamenti di informatica 2004/05.

Uso della tangente iperbolica

La tangente iperbolica $\tanh$ è anche nota sotto il nome di *funzione sigmoidea* e viene spesso utilizzata per modellare la crescita di popolazioni (*crescita logistica*) o la formazione di impulsi, ad esempio nelle *reti neuronali*. Come $\cosh$ e $\sinh$ si tratta di una funzione molto importante nelle applicazioni tecniche. La tangente iperbolica è definita come quoziente tra seno iperbolico e coseno iperbolico:

$$\tanh x := \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$$



Osserviamo che $\tanh 0 = 1$, $\lim_{x \to \infty} \tanh x = 1$, $\lim_{x \to -\infty} \tanh x = -1$.

Per $x \in \mathbb{R}^n$ siano $\mathrm{med}(x)$ la *mediana* di x e $x^{\mathrm{med}} := x - \mathrm{med}(x)^{\diamond}$. Si può adesso formare x^{med} e applicare ai componenti di x^{med} la funzione $\bigcirc_{\xi} \tanh \alpha\xi$ con un fattore α scelto in modo tale che i valori distinti nell'originale rimangano il più possibile distinguibili anche nel grafico trasformato. Per una matrice di dati X per ogni colonna X_i si dovrà scegliere un fattore α_i diverso. Con una notazione abbreviata la matrice trasformata diventa allora $(\tanh \alpha_1 X_1^{\mathrm{med}}, \ldots, \tanh \alpha_2 X_m^{\mathrm{med}})$.

La statistica del futuro

„The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently. Hyperspectral imagery, Internet portals, financial tick-by-tick data, and DNA microarrays are just a few of the better-known sources, feeding data in torrential streams into scientific and business databases ...

Classical methods are simply not designed to cope with this kind of explosive growth of dimensionality of the observation vector. We can say with complete confidence that in the coming century high-dimensional data analysis will be a very significant activity, and completely new methods of high-dimensional data analysis will be developed; we just don't know what they are." (David Donoho)

„È un momento particolarmente felice per la biostatistica in generale e per la statistica clinica in particolare. Gli sviluppi della biologia molecolare e della medicina stanno producendo enormi quantità di dati che devono essere ordinati e interpretati, creando così una domanda di competenze statistiche mai vista in precedenza.

Gli statistici clinici, cioè gli statistici che lavorano nella ricerca medica su umani, partecipano al clima di euforia grazie anche alla crescente disponibilità di risorse della ricerca medica, alla sua crescente matematizzazione e, con riferimento all'industria farmaceutica, grazie ad un ventennio di impressionante sviluppo. Per doveri legali e per tradizione culturale, l'industria farmaceutica è uno dei pochi settori produttivi che offre agli statistici la possibilità di una carriera non accademica di alto profilo scientifico. Insieme ai centri di cura e ricerca medica pubblici e privati, l'industria farmaceutica partecipa così attivamente alla richiesta e alla produzione di metodologia statistica." (Mauro Gasparini)

La struttura complessa e sorprendente degli spazi ad alta dimensione (trattati con più dettagli nelle pagine 1-5 del corso di Statistica multivariata 2005/06) crea difficoltà non solo in statistica, ma ad esempio anche negli *algoritmi di ricerca* in grandi insiemi di dati (basi di dati in medicina, nell'industria, in geografia, in biologia molecolare) che spesso vengono rappresentati (mediante tecniche sofisticate di trasformazione) come punti di qualche $\mathbb{R}_m$ ad alta e talvolta altissima dimensione. Gli algoritmi di ricerca classici spesso utilizzano concetti di *somiglianza* basati ad esempio sulla vicinanza nella metrica euclidea che però in questi spazi ad alta dimensione perde gran parte del suo significato. Superare questa difficiltà è uno dei compiti più attuali e più interessanti studiati dalla teoria delle basi di dati.

Ranghi

Molto utile in una prima fase dell'analisi è anche l'informazione sui *ranghi* dei valori nelle colonne della matrice X. Ciò in R avviene tramite la funzione `rank` di R che calcola per ogni elemento di un vettore il suo rango, cioè la posizione di quell'elemento nel vettore ordinato (che si ottiene con `sort`). Esempio:

```
x=c(3,5,1,10,9,2,8,6)
v=sort(x); print(v) # 1 2 3 5 6 8 9 10

u=rank(x)
print(u) # 3 4 1 8 7 2 6 5

x=c(2,5,6,2,1,3,5)
u=rank(x)
print(u) # 2.5 5.5 7 2.5 1 4 5.5
```

Come si vede nel secondo esempio, nell'impostazione iniziale, quando il vettore contiene valori uguali, `rank` assegna a questi elementi la media dei ranghi. Ciò nella visualizzazione grafica crea il problema che questi elementi (almeno in una dimensione) non sono più distinguibili. Con l'impostazione `ties.method='first'` in `rank` i ranghi diventano di nuovo unici, assegnando un rango minore a quelli tra due o più elementi uguali che nel vettore appaiono per primi:

```
x=c(2,5,6,2,1,3,5)
u=rank(x,ties.method='first')
print(u) # 2 5 7 3 1 4 6
```

Creiamo quindi due funzioni, di cui la seconda va applicata colonna per colonna a matrici di dati numerici, che calcolano i ranghi riportati a una scala che può essere impostata a seconda delle esigenze:

```
S.rango = function (x, scala=1)
{u=rank(x,ties.method='first')
(u-1)*scala/(length(x)-1)}

Sm.rango = function (X,scala=1)
apply(X,2,S.rango,scala=scala)
```

Nell'impostazione iniziale (`scala=1`) i ranghi vengono riportati all'intervallo $[0, 1]$, quindi per un vettore di 5 elementi otteniamo i punti 0, 0.25, 0.5, 0.75, 1:

```
x=c(2,8,1,3,2)
u=S.rango(x)
print(u) # 0.25 1 0 0.75 0.5
```

Visualizzazione di ranghi

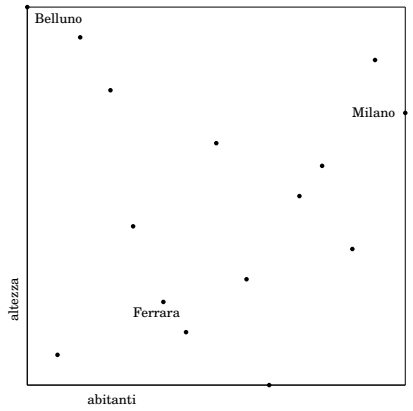
Se, per una figura su uno spazio di 50 mm, desideriamo in `S.rango` o `Sm.rango` una scala di 50, la possiamo reimpostare:

```
x=c(7,2,3,5,8,20,1,8,9,17,2)
u=S.rango(x,scala=50)
print(u) # 25 5 15 20 30 50 0 35 40 45 10
```

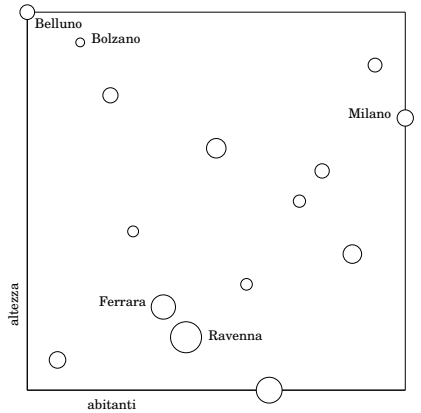
Queste funzioni si prestano particolarmente per la visualizzazione di matrici a due colonne. Applichiamo la funzione `Sm.rango` con `scala=50` alla matrice Y che consiste delle prime due colonne (relative al numero degli abitanti ed all'altezza sul mare) della matrice X che contiene i dati sui nostri 15 comuni. Otteniamo allora la matrice dei ranghi trasformati

$$\begin{pmatrix} 0 & 50 \\ 39 & 29 \\ 7 & 46 \\ 18 & 11 \\ 36 & 25 \\ 43 & 18 \\ 50 & 36 \\ 29 & 14 \\ 25 & 32 \\ 4 & 4 \\ 21 & 7 \\ 46 & 43 \\ 11 & 39 \\ 32 & 0 \\ 14 & 21 \end{pmatrix}$$

Riportiamo questi valori graficamente:



Se rappresentiamo ogni comune come cerchietto la cui area è proporzionale alla superficie, otteniamo



Lo studio dei ranghi è molto utile per una prima valutazione qualitativa delle relazioni tra le variabili; è però difficile tradurre questa visione qualitativa in criteri numerici che possano essere applicati successivamente ad altri insiemi di dati.

Correlazione di rango

La correlazione tra i vettori di rango si chiama la *correlazione di rango* di Spearman. Benché talvolta intuitiva e convincente, è difficile da interpretare numericamente.

Per il calcolo della correlazione di rango in genere si usano ranghi medi per elementi uguali la cui presenza crea qualche problema più teorico che pratico. In R la correlazione di rango tra due vettori x ed y la si ottiene quindi con

```
cor(rank(x),rank(y))
```

Usando `apply` possiamo anche calcolare la matrice delle correlazioni di rango per la matrice dei dati, ad esempio per i 15 comuni. Con

```
R=cor(apply(X,2,rank))
print(R)
```

otteniamo così

```
          ab    alt    mare   sup

ab     1     -0.06  -0.06   0.08
alt   -0.06   1      0.88  -0.54
mare  -0.06   0.88   1     -0.51
sup    0.08  -0.54  -0.51   1
```

dove abbiamo inserito a mano i titoli delle colonne e delle righe. La matrice è simmetrica perché lo è il coefficiente di correlazione ed è chiaro che la diagonale principale è occupata da 1. Vediamo tra l'altro che la correlazione (dei ranghi) tra il numero degli abitanti e le altre tre colonne è quasi zero, mentre la correlazione tra altezza e distanza dal mare è piuttosto alta (0.88).

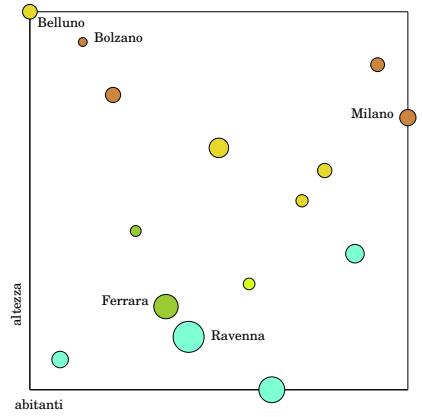
Colori e simboli

La rappresentazione grafica di dati ad alta dimensione è un importante strumento della statistica esploratoria. Essa può essere considerata una tecnica di trasformazione dei dati che però si appella nelle intenzioni d'utilizzo alle capacità della visione umana.

Nella seconda figura a pagina 28 siamo riusciti a rappresentare la superficie di ogni comune mediante l'area di un cerchio. Delle quattro variabili note sui 15 comuni ci manca soltanto la distanza dal mare che possiamo rappresentare almeno in modo qualitativo, tramite la scelta di un colore secondo il seguente schema in cui d denota la distanza dal mare in km:

$d \leq 12$ ... azzurro
$12 < d \leq 30$ ... verde chiaro
$30 < d \leq 60$ ... verde scuro
$60 < d \leq 100$ ... giallo scuro
$d > 100$ ... castano

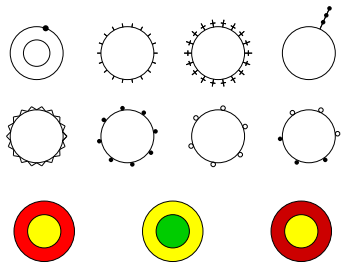
Ricordiamo che in ascisse e ordinata sono rappresentati i ranghi.



Il grafico a colori è già molto soddisfacente. Questa tecnica e le sue generalizzazioni possono essere utilizzate ogni volta che n, cioè il numero degli oggetti rappresentati, non è troppo grande in rapporto all'area disponibile; come in una buona mappa cartografica si possono rappresentare anche numerose variabili contemporaneamente. Se gli strumenti tipografici lo permettono si può scegliere un'area di disegno molto grande per poter aumentare n.

Nel disegno abbiamo rappresentato 15 oggetti su una superficie 5 cm $\times$ 5 cm; su una superficie di un m^2 possiamo quindi visualizzare $400 \times 15 = 6000$ oggetti, ad esempio i dati di 6000 pazienti.

Per rappresentare più variabili possiamo ornare i cerchietti e usare, se opportuno, più colori in ogni cerchietto:



In cartografia e meteorologia vengono utilizzati in modo sistematico numerosi simboli che permettono all'esperto di riconoscere facilmente situazioni anche molto complesse. Anche la notazione musicale può essere considerata come esempio di una sofisticata tecnica grafica per la rappresentazione di complicate serie temporali.

Quando il numero delle osservazioni è molto alto non è facile distinguerle in un diagramma bidimensionale; lo stesso vale, se ci sono osservazioni uguali. Allora si può usare un istogramma bidimensionale, che è quindi realizzato in 3 dimensioni e perciò a sua volta difficile da rappresentare e da interpretare in una grafica piana.

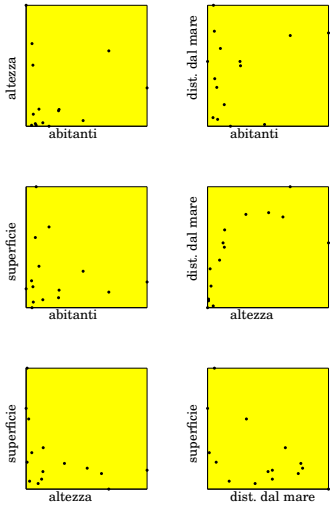
Spesso è da preferire l'uso di colori; la regione piana delle osservazione viene suddivisa in rettangoli della stessa forma e grandezza, e dopo la scelta di una scala di valori messa in corrispondenza con intervalli di frequenza ogni rettangolo viene colorato con il colore corrispondente al numero delle osservazioni in quel rettangolo.

Al posto di rettangoli si possono anche usare esagoni; si ottengono spesso grafici molto suggestivi, ma di non facile realizzazione sia per il disegno degli esagoni che per la corrispondenza dei colori.

Invece di una corrispondenza discreta si usano spesso, come peraltro per istogrammi univariati, funzioni continue per la rappresentazione di densità teoriche adattate alla situazione reale (*stime nucleari di densità*).

Rappresentazione a coppie

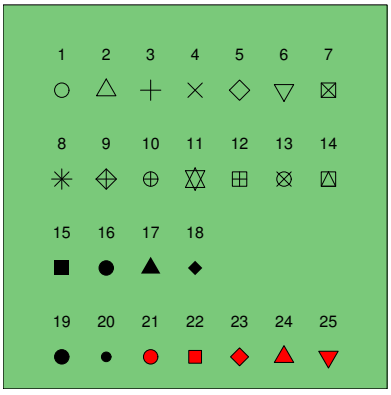
Più di due variabili possono essere confrontate a coppie come abbiamo fatto nella figura seguente per i 15 comuni. A differenza da pagina 28 qui abbiamo usato i valori numerici al posto dei ranghi.



Spesso si aggiungono anche i grafici riflessi all'asse dei 45 gradi; in R esiste una funzione apposita `pairs`. Il primo argomento è la matrice numerica dei dati, cioè la nostra X; per le opzioni grafiche possibili vedere `?pairs`. Provare

```
tab=read.table('Dati/Quindici-comuni')
X=Dt.matrice(tab)
pairs(X,pch=19)
locator(1)
dev.off()
```

Con il parametro `pch` si può scegliere la forma dei punti secondo questo schema:



L'immagine 2-dimensionale

Ci poniamo di nuovo nel contesto della situazione 19.1 La funzione `Smp` definita a pagina 25 calcola la matrice le cui colonne sono le componenti principali di X:

$$(X_{e^1}, \dots, X_{e^m})$$

La modifichiamo aggiungendo un secondo argomento facoltativo con cui possiamo calcolare determinate colonne di questa matrice:

```
Smp = function (X,j)
{XE=Smg.cen(X)%*%Smp.autovettori(X)
if (missing(j)) XE
else XE[,j]}
```

In questo modo con `Smp(X,1:2)` otteniamo le coordinate (cioè le lunghezze con segno rispetto al baricentro $\overline{X}$) delle proiezioni ortogonali dei punti X^i sul piano $\overline{X}+\mathbb{R}e^1+\mathbb{R}e^2$.

X sia la matrice dei dati per i 15 comuni introdotti a pagina 26. Per caricare i dati e per ottenere la matrice numerica X usiamo le istruzioni

```
Db(2)
X=Db.matrice()
```

Adesso con `XE=Smp(X)` otteniamo la matrice delle componenti principali e, siccome i coefficienti sono piuttosto grandi, li possiamo stampare con `print(round(XE,0))`:

```
-282 -216  215  49
  55  -48  -72 -11
-213 -252   76 -32
-215  187    5 -22
  53  -81  -92 -14
 296   83  -66  45
 973   29   23 -18
-113  -78 -140  24
-163   36  -14 -43
-239    1 -106  27
-228  420  114  -6
 580 -100   96   7
-215 -124   61 -26
 -78  257   17  19
-210 -114 -118  -1
```

Con

```
XE12=Smp(X,1:2)
print(round(XE12,0))
```

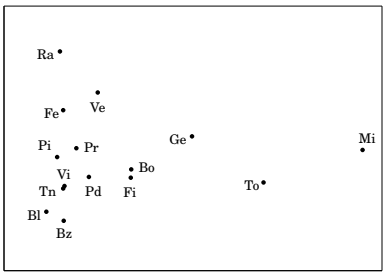
otteniamo quindi

```
-282 -216
  55  -48
-213 -252
-215  187
  53  -81
 296   83
 973   29
-113  -78
-163   36
-239    1
-228  420
 580 -100
-215 -124
 -78  257
-210 -114
```

La prima colonna è uguale ad X_{e^1}, la seconda ad X_{e^2}. Se riportiamo questi valori in un sistema cartesiano piano, otteniamo una proiezione

$$\mathbb{R}_4 \longrightarrow \mathbb{R}_2 \quad \text{con} \quad X^i \longmapsto (X^i_{e^1}, X^i_{e^2})$$

ottimale nel senso della nota 21.2.



Prima di ogni ragionamento matematico o statistico, proviamo a capire se questa proiezione può essere considerata convincente. E in effetti alcune configurazioni possono essere già intravviste: Milano si distingue fortemente dagli altri comuni, e i comuni più vicini sono le altre grandi città, soprattutto Torino e Genova e poi Bologna e Firenze. Non è un caso che andiamo verso sinistra perché è appunto l'asse orizzontale quello con la varianza maggiore. Vediamo che seguono verso sinistra Venezia, Parma e Padova, e poi gli altri comuni, con quelli più vicini al mare (in particolare Ravenna e Venezia) più in alto, e le città di montagna (Bolzano, Belluno, Trento) più in basso. La rappresentazione 2-dimensionale che abbiamo ottenuto dalle componenti principali è quindi già piuttosto soddisfacente.

Per valutare l'affidabilità matematica calcoliamo, con `Smp.autovalori(X)`, gli autovalori di Ω, ottenendo dopo arrotondamento i valori

$$\lambda_1 = 1791717 \qquad \lambda_2 = 450423$$
$$\lambda_3 = 141728 \qquad \lambda_4 = 10903$$

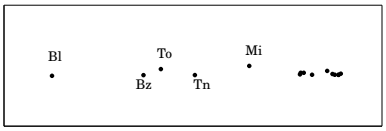
per cui

$$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = \frac{1791717+450423}{1791717+450423+141728+10903} = \frac{2242140}{2394771} = 0.936$$

Perché bisogna standardizzare

Siccome le componenti principali dipendono fortemente dalle scale di misura usate per le variabili, i dati devono sempre essere standardizzati, usando $\widehat{X}$, X^{01}, la matrice dei ranghi o un'altra trasformazione per ottenere una forma dei dati che possiede opportune proprietà di invarianza.

Assumiamo di aver misurato le altezze dei 15 comuni in centimetri. Allora nella matrice dei dati la seconda colonna deve essere moltiplicata per 100. Procedendo con la matrice così ottenuta come nella prima figura, otteniamo la seguente figura:



Si vede chiaramente che l'altezza determina in pratica da sola la proiezione cancellando quasi del tutto il significato delle altre variabili. Il rapporto di variazione stavolta è addirittura uguale a 0.9998 ma ciò, come si vede, non garantisce un risultato soddisfacente.

È quindi sempre necessario effettuare una standardizzazione. In alcuni casi ci possono essere ragioni per attribuire *pesi* diversi alle variabili, lavorando ad esempio con $(\widehat{X}_1, 2\widehat{X}_2, 0.4\widehat{X}_3)$, se la seconda variabile ci sembre più importante della prima e questa a sua volta più importante della terza. Una tale scelta deve però essere giustificata dalle caratteristiche dei dati.

Se più colonne della matrice dei dati esprimono lo stesso fenomeno, esse naturalmente avranno più peso in un'analisi delle componenti principali e questa molteplicità di colonne essenzialmente uguali non è eliminata dalle standardizzazioni finora viste. Ciò mostra che è molto importante pianificare in anticipo quali variabili vogliamo scegliere per l'analisi statistica. Talvolta anche qui può aiutare l'analisi delle componenti principali di X^t.

Il matematico in statistica

Per fare bene il suo lavoro, lo statistico che lavora in un'azienda, nell'amministrazione pubblica o nella ricerca clinica, deve comprendere i compiti che gli vengono posti e deve essere in grado di interagire con i committenti. Nonostante ciò la statistica è di sua natura una disciplina matematica che si basa sul calcolo delle probabilità, una teoria astratta e difficile, e richiede conoscenze tecniche in altri campi della matematica come analisi reale e complessa, analisi armonica, calcolo combinatorio (ad esempio per la pianificazione di esperimenti). Nell'analisi delle componenti principali e nella ricerca di raggruppamenti sarà compito dello statistico scegliere la rappresentazione dei dati e le misure per la somiglianza o diversità di individui e gruppi. In questo corso abbiamo potuto accennare solo ad alcune delle difficoltà concettuali e tecniche che si incontrano.

Nella statistica multivariata in particolare probabilmente molte tecniche sono ancora da scoprire e i metodi più efficienti si baseranno forse su metodi geometrici avanzati, ad esempio della geometria algebrica reale e della teoria delle rappresentazioni di gruppi.

Ci sono tanti campi di applicazione della statistica in medicina, bioinformatica, farmacologia, matematica finanziaria, linguistica, demografia, che uno studente che intraprende questa professione dopo aver acquisito una solida formazione matematica può sperare in un'attività interessante e gratificante.

L'abitudine ai dati e alla loro interpretazione formerà le sue capacità di giudicare situazioni complesse in modo razionale oltre a fornirgli un ricco patrimonio di informazioni, quindi potrà anche aspirare a una carriera amministrativa o manageriale.

Nel suo lavoro giornaliero potrà, nei contatti con ricercatori clinici o amministratori o con l'opinione pubblica utilizzare le proprie conoscenze teoriche per chiarire il significato di risultati di test clinici o di rilievi statistici o per proporre nuovi esperimenti o indagini.

La standardizzazione $\widehat{X}$

Definizione 31.1. Poniamo

$$\widehat{X} := (\widehat{X_1}, \dots, \widehat{X_m})$$

Sappiamo dall'osservazione 3.6 che

$$\overline{\widehat{X_j}} = 0$$

per ogni j e quindi

$$C\widehat{X} = \widehat{X}$$

Ciò a sua volta implica che

$$\Omega_{\widehat{X}} = \widehat{X}^t \widehat{X}$$

Sostituendo la matrice X con $\widehat{X}$, possiamo perciò applicare la teoria finora sviluppata a questa nuova matrice.

Poniamo $Y := \widehat{X}$. Allora

$$(Y^t Y)^i_j = \|\widehat{X_i}, \widehat{X_j}\| = r_{X_i X_j}$$

Per questa ragione la matrice $\Omega_{\widehat{X}}$ si chiama anche la *matrice di correlazione* di X. Essa in R può essere ottenuta semplicemente con `cor(X)`. Per trovare $\widehat{X}$ possiamo definire la funzione

```
Smg.ng = function (X)
apply(X,2,Sg.ng)
```

Sia adesso X la matrice dei 15 comuni; definiamo $Y := \widehat{X}$ come sopra e procediamo come a pagina 30, sostituendo X con Y.

```
Db(2)
X=Db.matrice()
Y=Smg.ng(X)
print(round(Y,2))
```

ottenendo prima $Y = \widehat{X}$:

```
-0.22  0.65  0.08 -0.12
 0.04 -0.10  0.05 -0.13
-0.17  0.37  0.47 -0.27
-0.15 -0.20 -0.10  0.30
 0.03 -0.10  0.08 -0.19
 0.23 -0.17 -0.35  0.03
 0.73  0.06  0.28 -0.06
-0.09 -0.19 -0.22 -0.20
-0.12 -0.09  0.17  0.07
-0.18 -0.21 -0.30 -0.05
-0.14 -0.21 -0.32  0.71
 0.43  0.32  0.26 -0.14
-0.17  0.22  0.29 -0.10
-0.04 -0.22 -0.36  0.38
-0.17 -0.13 -0.04 -0.22
```

A questo punto con

```
YE=Smp(Y)
print(round(YE,2))
```

calcoliamo le componenti principali di Y:

```
-0.43 -0.39  0.24  0.31
-0.05  0.05 -0.14 -0.06
-0.60 -0.29  0.05 -0.11
 0.35 -0.07  0.11 -0.13
-0.10  0.05 -0.19 -0.08
 0.28  0.27 -0.12  0.20
-0.37  0.68  0.10 -0.05
 0.15 -0.05 -0.32  0.08
 0.00 -0.09  0.03 -0.22
 0.31 -0.12 -0.23  0.09
 0.71 -0.04  0.40 -0.05
-0.50  0.32  0.14  0.09
-0.32 -0.23  0.07 -0.10
 0.55  0.04  0.13  0.07
 0.01 -0.14 -0.27 -0.03
```

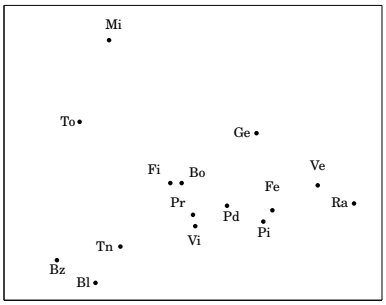
oppure, con

```
YE12=Smp(Y,1:2)
print(round(YE12,2))
```

le prime due componenti principali di Y:

```
-0.43 -0.39
-0.05  0.05
-0.60 -0.29
 0.35 -0.07
-0.10  0.05
 0.28  0.27
-0.37  0.68
 0.15 -0.05
 0.00 -0.09
 0.31 -0.12
 0.71 -0.04
-0.50  0.32
-0.32 -0.23
 0.55  0.04
 0.01 -0.14
```

Riportiamo anche stavolta questi valori in un sistema cartesiano piano:



Sicuramente la risoluzione in questo caso è migliore che prima della standardizzazione; anche i gruppi che possiamo formare, ad esempio

Milano, Torino
Genova, Venezia, Ravenna
Ferrara, Padova, Pisa
Firenze, Bologna, Parma, Vicenza
Trento, Bolzano, Belluno

sono abbastanza convincenti. Forso l'unico dubbio potrebbe riguardare la vicinanza tra Ferrara e Pisa (bisogna però anche tener conto dei dati che avevamo a disposizione) e la notevole distanza tra Trento e Vicenza molto vicine nella prima proiezione. Calcoliamo anche qui gli autovalori con `Smp.autovalori(Y)`, ottenendo

$$\lambda_1 = 2.18$$
$$\lambda_2 = 0.98$$
$$\lambda_3 = 0.58$$
$$\lambda_4 = 0.26$$

Nonostante la favorevole impressione, stavolta il secondo rapporto di variazione è uguale a 0.789 e perciò più basso di quello ottenuto a pagina 30; ma siamo partiti da standardizzazioni diverse. Anche qui, come quando si osserva un oggetto tridimensionale in natura, è utile osservarlo da prospettive diverse.

Esercizio 31.2. Definendo

$$\check{X} := (\check{X_1}, \dots, \check{X_m})$$

si ha

$$\check{X}^t \check{X} = (n-1) \cdot \widehat{X}^t \widehat{X}$$

Se definiamo quindi $Y := \widehat{X}$, $Z := \check{X}$, allora le componenti principali di Z si distinguono da quelle di Y solo per un fattore $\sqrt{n-1}$, per cui otteniamo risultati sostanzialmente equivalenti.

La standardizzazione X^{01}

Proviamo adesso ad applicare il metodo generale alla matrice X^{01} che si ottiene da X mediante proiezione su $[0, 1]$. Per i quindici comuni X^{01} è già stata calcolata a pagina 27. Con

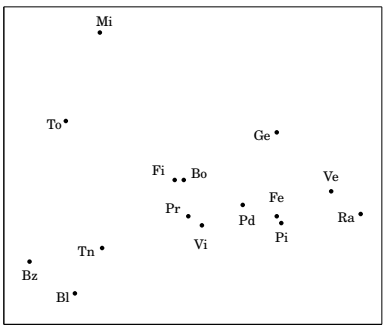
```
Db(2)
X=Db.matrice()
X01=Sm.tra01(X)

X01E=Smp(X01)
print(round(X01E,2))

X01E12=Smp(X01,1:2)
print(round(X01E12,2))
```

otteniamo la matrice delle componenti principali

```
-0.52 -0.43  0.24  0.35
-0.04  0.07 -0.15 -0.07
-0.72 -0.29  0.01 -0.13
 0.37 -0.09  0.12 -0.16
-0.08  0.07 -0.21 -0.09
 0.37  0.28 -0.09  0.23
-0.41  0.72  0.12 -0.05
 0.22 -0.04 -0.33  0.10
-0.02 -0.09  0.01 -0.26
 0.39 -0.12 -0.23  0.11
 0.74 -0.08  0.44 -0.08
-0.56  0.33  0.15  0.11
-0.40 -0.23  0.05 -0.12
 0.61  0.02  0.16  0.07
 0.04 -0.13 -0.30 -0.03
```

e le prime due colonne, che corrispondono alle prime due componenti principali di X^{01} e che poi rappresentiamo in $\mathbb{R}_2$:

```
-0.52 -0.43
-0.04  0.07
-0.72 -0.29
 0.37 -0.09
-0.08  0.07
 0.37  0.28
-0.41  0.72
 0.22 -0.04
-0.02 -0.09
 0.39 -0.12
 0.74 -0.08
-0.56  0.33
-0.40 -0.23
 0.61  0.02
 0.04 -0.13
```



Il risultato è molto simile a quello ottenuto per $\widehat{X}$. Anche il rapporto di variazione 0.79 è praticamente identico.

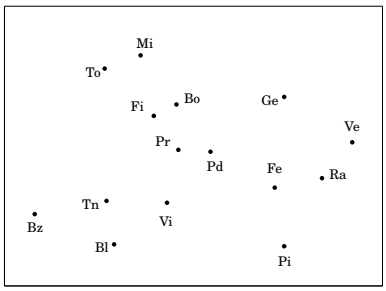
Analisi della matrice dei ranghi

Eseguiamo infine l'analisi delle componenti principali per la matrice dei ranghi. Con

```
Db(2); X=Db.matrice()
XR=Sm.rango(X)
XRE12=Smp(XR,1:2)
print(round(XRE12,2))
```

otteniamo le prime due componenti principali, che riportiamo nel piano:

```
-0.42 -0.46
-0.09  0.28
-0.84 -0.30
 0.43 -0.16
-0.21  0.22
 0.48  0.32
-0.28  0.54
 0.09  0.03
-0.08  0.04
 0.48 -0.47
 0.68 -0.11
-0.47  0.47
-0.46 -0.23
 0.84  0.08
-0.14 -0.24
```



La risoluzione è molto buona e la classificazione in gruppi convincente. Anche qui vediamo che l'uso dei ranghi introduce degli aspetti che sfuggono talvolta all'analisi puramente metrica-lineare. Gli autovalori sono $\lambda_1 = 3.28, \lambda_2 = 1.42, \lambda_3 = 0.81, \lambda_4 = 0.19$, il rapporto di variazione è 0.82.

screeplot

Combinando `princomp` con `screeplot` si possono visualizzare i rapporti tra gli autovalori. Provare dal terminale

```
Db(2); X=Db.matrice(); p=princomp(X)
screeplot(p)
screeplot(p,type='lines')
summary(p)
```

Analisi di X^t

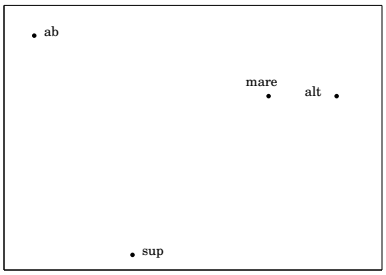
Molto spesso può essere utile studiare anche la trasposta X^t della matrice dei dati mediante un'analisi delle componenti principali. Usiamo la proiezione su $[0, 1]$ come standardizzazione e procediamo come a pagina 34:

```
Db(2)
X=Db.matrice()
Xt=t(X)
Xt01=Sm.tra01(Xt)

CP12=Smp(Xt01,1:2)
print(round(CP12,2))
```

ottenendo così le prime due delle 15 componenti principali:

```
-1.64  0.88
 1.47  0.24
 0.80  0.26
-0.63 -1.39
```

che possiamo riportare anche in questo caso in un sistema cartesiano:

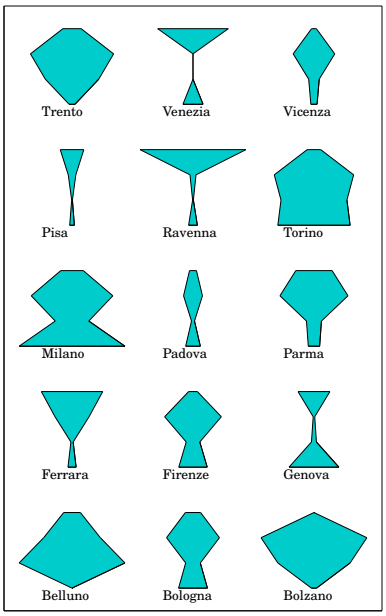


La figura, una proiezione 2-dimensionale di $\mathbb{R}^{15}$, mostra la vicinanza tra i fattori *distanza dal mare* e *altezza*. In un'indagine medica, dove le colonne corrispondono a caratteristiche cliniche e le righe a pazienti oppure le righe a cellule tumorali e le colonne a geni di ciascuno dei quali per ogni cellula è indicata l'intensità di espressione, in questo modo si possono individuare gruppi di fattori o geni con effetti vicini. Una discussione di tecniche multivariate nello studio di microarray di DNA si trova nel libro di Lee.

M. Lee: Analysis of microarray gene expression data. Kluwer 2004.

Biprofili

Quando sia il numero delle variabili che quello degli oggetti non sono troppo grandi, spesso si ottengono risultati interessanti con i *biprofili* (figure a vaso) che consistono semplicemente nel raddoppio del grafico dei valori spesso riportato rispetto a un'ascisse verticale. Per i nostri 15 comuni, applicando la funzione `Sm.tra01` alla matrice X, otteniamo le rappresentazioni



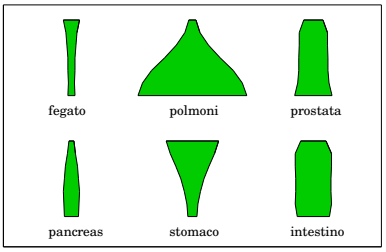
Queste figure possono però anche ingannare se non si osservano esattamente i confini tra le variabili indicate; in particolare un valore grande di una variabile crea un'area grande in corrispondenza ad essa che otticamente sembra venga condivisa con le variabili adiacenti anche quando queste hanno valori piccoli. Sono comunque ben visibili somiglianze tra Firenze e Bologna, tra Bolzano e Trento, tra Genova e Venezia.

Le figure a vaso dipendono naturalmente dall'ordine in cui gli elementi dei vettori di dati sono elencati; quando esiste un ordine naturale, ad esempio in una serie temporale o in statistiche demografiche, anche l'interpretazione della figura è più naturale.

Rappresentiamo in questo modo una statistica di mortalità per diversi tipi di cancro nella popolazione statunitense maschile (da De Vita/, pag. 233), in con le seguenti abbreviazioni:

Pa ... pancreas	St ... stomaco
In ... intestino	Fe ... fegato
Po ... polmoni	Pr ... prostata

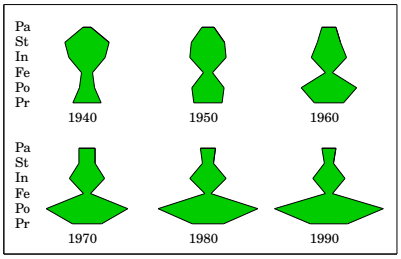
	Pa	St	In	Fe	Po	Pr
1930	5	52	25	16	7	19
1940	8	44	36	11	15	28
1950	12	33	36	9	32	28
1960	14	23	35	7	55	29
1970	16	16	35	7	81	29
1980	15	11	36	6	99	31
1990	14	9	32	7	108	37



Il 1990 si trova alla parte inferiore dei vasi.

Si nota un fortissimo aumento della mortalità per cancro ai polmoni e una notevole diminuzione per il cancro allo stomaco, mentre i valori per gli altri tipi di cancro sono variati molto meno nei 60 anni tra il 1930 e il 1990, con un aumento a più del doppio comunque nel cancro al pancreas.

Interessante è anche la distribuzione dei tipi di cancro nei singoli anni, da cui vediamo che mentre nel 1940 i cancri allo stomaco e all'intestino causavano più vittime, negli anni più recenti gli organi più colpiti sono polmone e, in misura minore, la prostata.



V. De Vita/S. Hellman/S. Rosenberg (ed.): Cancer. Lippincott-Raven 1997.

VII. REGRESSIONE MULTIVARIATA

Regressione semplice in forma matriciale

Osservazione 33.1. *Siano $X_1, \dots, X_k \in \mathbb{R}^n$ ed $X := (X_1, \dots, X_k)$. Allora* $\mathrm{SV}(X_1, \dots, X_k) = \{Xa \mid a \in \mathbb{R}^k\}$.

Osservazione 33.2. *Sia $A \in \mathbb{R}^n_m$. La matrice quadratica A^tA è invertibile se e solo se le colonne di A sono linearmente indipendenti.*

Dimostrazione. L'enunciato è in pratica già contenuto nell'osservazione 15.2 o nel corollario 15.3. Diamo una dimostrazione diretta utilizzando solo l'osservazione 15.1.

(1) A^tA sia invertibile. Se le colonne di A sono linearmente dipendenti, allora esiste $u \in \mathbb{R}^m \setminus 0$ con $Au = 0$. Ma allora $A^tAu = 0$ e ciò implica $u = 0$ perché A^tA è invertibile. Abbiamo una contraddizione.

(2) Le colonne di A siano linearmente indipendenti, ma A^tA non sia invertibile. Allora esiste $u \in \mathbb{R}^m \setminus 0$ con $A^tAu = 0$. Ma allora anche $u^tA^tAu = 0$, cioè $\|Au, Au\| = 0$ e quindi $Au = 0$. Ciò per ipotesi implica $u = 0$, una contraddizione.

Osservazione 33.3. *La matrice* $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *sia invertibile.*

Allora $A^{-1} = \dfrac{1}{\det A}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Nota 33.4. Come nella situazione 3.3 siano $x, y \in \mathbb{R}^n$ due vettori non diagonali. Formiamo la matrice $X := (x, 1^\diamond) \in \mathbb{R}^n_2$. Per ipotesi le due colonne di X sono linearmente indipendenti. Come nella nota 4.2 siano P_x il piano generato da x e $1^\diamond$ e p la proiezione di y su P_x. Come in precedenza, una volta trovato p, denotiamo con λ e τ i coefficienti nella rappresentazione $p = \lambda x + \tau^\diamond$.

Per l'osservazione 33.1 $P_x = \{Xa \mid a \in \mathbb{R}^2\}$. Dobbiamo quindi cercare $a \in \mathbb{R}^2$ in modo tale che $p = Xa$ ed $y - Xa \perp P_x$. Questa condizione è equivalente a $X^t(y - Xa) = 0$, cioè $X^ty = X^tXa$.

Per l'osservazione 33.2 la matrice quadratica X^tX è invertibile e quindi abbiamo $a = (X^tX)^{-1}X^ty$, cosicché $p = X(X^tX)^{-1}X^ty$.

Dimostriamo che $a = \begin{pmatrix} \lambda \\ \tau \end{pmatrix}$ con λ, τ come nella nota 4.2. Abbiamo infatti

$$X = \begin{pmatrix} x^1 & 1 \\ \vdots & \vdots \\ x^n & 1 \end{pmatrix} \quad \text{e} \quad X^t = \begin{pmatrix} x^1 & \dots & x^n \\ 1 & \dots & 1 \end{pmatrix}$$

$$X^tX = X^t = \begin{pmatrix} x^1 & \dots & x^n \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} x^1 & 1 \\ \vdots & \vdots \\ x^n & 1 \end{pmatrix} = \begin{pmatrix} |x|^2 & n\overline{x} \\ n\overline{x} & n \end{pmatrix}$$

e

$$\det X^tX = n|x|^2 - n^2\overline{x}^2 = n|Cx|^2$$

Per l'osservazione 33.3 allora

$$(X^tX)^{-1} = \frac{1}{n|Cx|^2}\begin{pmatrix} n & -n\overline{x} \\ -n\overline{x} & |x|^2 \end{pmatrix}$$

Inoltre

$$X^ty = \begin{pmatrix} x^1 & \dots & x^n \\ 1 & \dots & 1 \end{pmatrix}\begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix} = \begin{pmatrix} \|x, y\| \\ n\overline{y} \end{pmatrix}$$

per cui

$$\begin{aligned}(X^tX)^{-1}X^ty &= \frac{1}{n|Cx|^2}\begin{pmatrix} n & -n\overline{x} \\ -n\overline{x} & |x|^2 \end{pmatrix}\begin{pmatrix} \|x, y\| \\ n\overline{y} \end{pmatrix} \\ &= \frac{1}{n|Cx|^2}\begin{pmatrix} n\|x, y\| - n^2\overline{x}\,\overline{y} \\ -n\overline{x}\|x, y\| + |x|^2 n\overline{y} \end{pmatrix} \\ &= \frac{1}{|Cx|^2}\begin{pmatrix} \|x, y\| - n\overline{x}\,\overline{y} \\ |x|^2\overline{y} - \overline{x}\|x, y\| \end{pmatrix} = \begin{pmatrix} a^1 \\ a^2 \end{pmatrix}\end{aligned}$$

Vediamo direttamente che

$$a^1 = \frac{\|x, y\| - n\overline{xy}}{|Cx|^2} = \frac{\|Cx, Cy\|}{|Cx|^2} = \lambda$$

mentre

$$\begin{aligned}a^2 &= \frac{|x|^2 - n\overline{x}^2}{|x|^2 - n\overline{x}^2}\overline{y} + \frac{n\overline{x}^2\overline{y} - \overline{x}\|x, y\|}{|x|^2 - n\overline{x}^2} \\ &= \overline{y} + \frac{n\overline{x}\,\overline{y} - \|x, y\|}{|x|^2 - n\overline{x}^2}\overline{x} = \overline{y} - \lambda\overline{x} = \tau\end{aligned}$$

Proposizione 33.5. *Per $a, b, c, d \in \mathbb{R}^n$ vale la relazione*

$$\|a \wedge b, c \wedge d\| = \begin{vmatrix} \|a, c\| & \|a, d\| \\ \|b, c\| & \|b, d\| \end{vmatrix}$$

Nel caso $n = 3$ invece del prodotto esterno possiamo utilizzare anche il prodotto vettoriale:

$$\|a \times b, c \times d\| = \begin{vmatrix} \|a, c\| & \|a, d\| \\ \|b, c\| & \|b, d\| \end{vmatrix}$$

Dimostrazione. Corso di Algoritmi e strutture di dati 2006/07, pag. 32, per il caso $n = 3$; corsi di Geometria per il caso generale.

Nota 33.6. Nei conti della nota 33.4 non abbiamo usato in modo essenziale che la seconda colonna di X era il vettore $1^\diamond$, ma solo che le due colonne siano linearmente indipendenti, in modo che la matrice X^tX sia invertibile.

Siano quindi dati due vettori linearmente indipendenti X_1, X_2 di $\mathbb{R}^n$ e sia $X := (X_1, X_2)$. Possiamo allora risolvere il problema di minimizzare $|y - \lambda^1X_1 - \lambda^2X_2|$ o equivalentemente $|y - Xa|$ per $a \in \mathbb{R}^2$.

Ciò geometricamente è equivalente alla condizione che $y - Xa$ sia ortogonale al piano $\mathrm{SV}(X_1, X_2)$ e quindi a $X^t(y - Xa) = 0$, ovvero

$$X^ty = X^tXa \quad (\textit{equazione normale})$$

Le ipotesi e l'osservazione 33.2 implicano che $a = (X^tX)^{-1}X^ty$.

Rifacendo i conti in questo caso più generale abbiamo X^tX

$$= \begin{pmatrix} X_1^1 & \dots & X_1^n \\ X_2^1 & \dots & X_2^n \end{pmatrix}\begin{pmatrix} X_1^1 & X_2^1 \\ \vdots & \vdots \\ X_1^n & X_2^n \end{pmatrix} = \begin{pmatrix} \|X_1, X_1\| & \|X_1, X_2\| \\ \|X_1, X_2\| & \|X_2, X_2\| \end{pmatrix}$$

per cui, con $\Delta := \|X_1, X_1\|\|X_2, X_2\| - \|X_1, X_2\|^2$,

$$(X^tX)^{-1} = \frac{1}{\Delta}\begin{pmatrix} \|X_2, X_2\| & -\|X_1, X_2\| \\ -\|X_1, X_2\| & \|X_1, X_1\| \end{pmatrix}$$

mentre

$$X^ty = \begin{pmatrix} X_1^1 & \dots & X_1^n \\ X_2^1 & \dots & X_2^n \end{pmatrix}\begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix} = \begin{pmatrix} \|y, X_1\| \\ \|y, X_2\| \end{pmatrix}$$

per cui $a = \begin{pmatrix} \lambda^1 \\ \lambda^2 \end{pmatrix} = \dfrac{1}{\Delta}\begin{pmatrix} \|X_2, X_2\|\|y, X_1\| - \|X_1, X_2\|\|y, X_2\| \\ -\|X_1, X_2\|\|y, X_1\| + \|X_1, X_1\|\|y, X_2\| \end{pmatrix}$

Dalla proposizione 33.5 vediamo però che ciò può essere scritto anche nella forma

$$\lambda^1 = \frac{\|X_1 \wedge X_2, y \wedge X_2\|}{\|X_1 \wedge X_2, X_1 \wedge X_2\|}$$

$$\lambda^2 = \frac{\|X_1 \wedge X_2, X_1 \wedge y\|}{\|X_1 \wedge X_2, X_1 \wedge X_2\|}$$

Ciò ci induce alla congettura che per m arbitrario avremmo, con notazione analoga,

$$\lambda^i = \frac{\|X_1 \wedge \dots \wedge X_m, X_1 \wedge \dots \wedge X_{i-1} \wedge y \wedge X_{i+1} \wedge \dots \wedge X_m\|}{\|X_1 \wedge \dots \wedge X_m, X_1 \wedge \dots \wedge X_m\|}$$

per ogni i.

E infatti è così come dimostreremo adesso utilizzando la regola di Cramer.

Regressione lineare multivariata

Proposizione 34.1. *Siano $u_1, \dots, u_m, v_1, \dots, v_m \in \mathbb{R}^n$. Allora*

$$\|u_1 \wedge \dots \wedge u_m, v_1 \wedge \dots \wedge v_m\| = \begin{vmatrix} \|u_1, v_1\| & \dots & \|u_1, v_m\| \\ & \dots & \\ \|u_m, v_1\| & \dots & \|u_m, v_m\| \end{vmatrix}$$

Dimostrazione. Corsi di Geometria.

Nota 34.2. Consideriamo ancora il caso generale della regressione lineare multivariata. Siano $y \in \mathbb{R}^n$ ed $X \in \mathbb{R}^n_m$. Le colonne di X siano linearmente indipendenti. Cerchiamo la proiezione p di y sull'iperpiano $\mathrm{SV}(X_1, \dots, X_m)$ generato da $X_1, \dots, X_m$.

Anche in questo caso p può essere scritto nella forma $p = Xa$ con $a \in \mathbb{R}^m$. Come nella nota 33.6 dalla condizione di ortogonalità $y - Xa \perp \mathrm{SV}(X_1, \dots, X_m)$ segue l'equazione normale

$$X^t(y - Xa) = 0 \text{ ovvero } X^t y = X^t X a$$

Per ipotesi le colonne di X sono linearmente indipendenti, perciò la matrice $X^t X$ è invertibile cosicché troviamo

$$a = (X^t X)^{-1} X^t \text{ e quindi } p = X(X^t X)^{-1} X y$$

Se scriviamo $a = \begin{pmatrix} \lambda^1 \\ \vdots \\ \lambda^m \end{pmatrix}$, dalla regola di Cramer abbiamo

$$\lambda^i = \frac{\det((X^t X)_1, \dots, (X^t X)_{i-1}, X^t y, (X^t X)_{i+1}, \dots, (X^t X)_m)}{\det X^t X}$$

per ogni i. Però $(X^t X)_k = X^t X_k$ per ogni k, cosicché il vettore $(X^t X)_k$ può essere scritto nella forma $\begin{pmatrix} \|X_1, X_k\| \\ \vdots \\ \|X_m, X_k\| \end{pmatrix}$.

Dalla proposizione 34.1 otteniamo adesso

$$\lambda^i = \frac{\|X_1 \wedge \dots \wedge X_m, X_1 \wedge \dots \wedge X_{i-1} \wedge y \wedge X_{i+1} \wedge \dots \wedge X_m\|}{\|X_1 \wedge \dots \wedge X_m, X_1 \wedge \dots \wedge X_m\|}$$

in accordo con la congettura formulata alla fine della nota 33.6.

Osservazione 34.3. *Anche nel caso generale della nota 34.2 si lavora spesso con un'ultima colonna $X_m = 1^\diamond$. Ciò significa che si considerano $m-1$ variabili indipendenti e si cerca un'approssimazione ottimale di y della forma*

$$y \sim \lambda^1 X_1 + \dots + \lambda^{m-1} X_{m-1} + \tau^\diamond$$

Osservazione 34.4. In R i coefficienti di una regressione lineare multivariata si trovano con istruzioni della forma

```
coeff = lm (y ~ X1 + X2 + ... + Xm)
```

Esempio 34.5. Consideriamo di nuovo i 15 comuni. Vogliamo verificare se il numero di abitanti di un comune possa essere espresso approssimativamente in modo lineare mediante le variabili altezza, distanza dal mare e superficie.

Dopo aver caricato e definito i dati con

```
Db(2); A=Db.matrice()
ab=A[,1]; alt=A[,2]; mare=A[,3]; sup=A[,4]
```

con

```
coeff=lm(ab~alt+mare+sup); print(coeff)
```

troviamo questi coefficienti:

```
# (Intercept)          alt         mare          sup
#   221.85265     -0.66857      2.93906     -0.03075
```

Il modello di regressione è quindi

```
ab ~ -0.67 alt + 2.94 mare -0.03 sup + 221.85
```

Vediamo in particolare che nel nostro esempio il numero degli abitanti di un comune non dipende dalla superficie. Per confrontare il modello con i dati originali usiamo le seguenti istruzioni:

```
X=A[,2:4]
f=function (r) 221.85-0.67*r[1]+2.94*r[2]-0.03*r[3]

abr=apply(X,1,f)

confronto=matrix(c(ab,abr),ncol=2); print(confronto)
# Output:

   35 181.30
  380 387.24
   97 456.32
        ...
```

È evidente che i valori previsti dalla regressione multivariata corrispondono solo in pochissimi casi al vero numero di abitanti. Nel nostro esempio quindi il modello lineare è del tutto inadeguato. Se esiste una qualche relazione causale tra le variabili (e non è detto che esista) sicuramente non può essere lineare.

Esercizio: Per la superficie si trova il modello lineare

```
sup ~ 346.17 - 0.12 alt - 1.84 mare
```

Confrontare i valori che si ottengono con i dati originali. Stavolta la corrispondenza è (con due eccezioni) molto migliore. Il modello in pratica prevede superficie grandi per comuni vicini al mare e ciò almeno in parte corrisponde alla realtà.

Osservazione 34.6. In uno studio concreto bisogna quindi prima verificare se un modello lineare può essere appropriato. Stabilito ciò (con appositi strumenti statistici) si può tentare di interpretare i coefficienti. Per poterli confrontare bisogna sottoporre la matrice (y, X) a una standardizzazione, ad esempio sostituendo y con $\widehat{y}$ ed X con $\widehat{X}$. Ciò può però soltanto contribuire a un'interpretazione dei rapporti di grandezza tra i coefficienti di regressione, ma non migliora la qualità del modello.

Osservazione 34.7. Anche nella valutazione e nella scelta di un modello di regressione multivariata le rappresentazioni grafiche sono molto importanti. Da esse si può spesso indovinare una trasformazione dei dati che li rende suscettibili a un modello lineare o polinomiale.

G. Seber/A. Lee: Linear regression analysis. Wiley 2003.
G. Seber/C. Wild: Nonlinear regression. Wiley 2003.

Regressione polinomiale

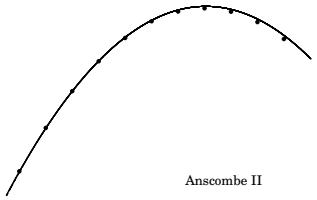
Nota 34.8. La regressione lineare multivariata in verità comprende anche la regressione polinomiale. Infatti un modello della forma

$$y^i = \alpha + \beta x^i + \gamma (x^i)^2$$

può essere considerato come un problema di regressione lineare in cui la colonna X_2 consiste dei quadrati dei valori nella colonna X_1. Applichiamo questa idea al secondo esempio di Anscombe (pagina 10), in cui si ha la netta impressione che y dipenda in modo quadratico da x:

```
y=c(9.14,8.14,8.74,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74)
X1=c(10,8,13,9,11,14,6,4,12,7,5); X2=X1^2
X=matrix(c(X1,X2),ncol=2)

coeff=lm(y ~ X1 + X2); print(coeff)

# Output: -5.9957  2.7808  -0.126
```

Con la parabola $y = -6 + 2.78x - 0.126x^2$ i dati vengono approssimati molto bene:



Anscombe II

VIII. OTTIMIZZAZIONE GENETICA

Problemi di ottimizzazione

Siano dati un insieme X, un sottoinsieme A di X e una funzione $f : X \longrightarrow \mathbb{R}$. Cerchiamo il minimo di f su A, cerchiamo cioè un punto $a_0 \in A$ tale che $f(a_0) \leq f(a)$ per ogni $a \in A$. Ovviamente il massimo di f è il minimo di $-f$, quindi vediamo che non è una restrizione se in seguito in genere parliamo solo di uno dei due.

Ci si chiede a cosa serve l'insieme X, se il minimo lo cerchiamo solo in A. La ragione è che spesso la funzione è data in modo naturale su un insieme X, mentre A è una parte di X descritta da condizioni aggiuntive. Quindi i punti di X sono tutti quelli in qualche modo considerati, i punti di A quelli *ammissibili*. In alcuni casi le condizioni aggiuntive (dette anche *vincoli*) non permettono di risalire facilmente ad A, e può addirittura succedere che la parte più difficile del problema sia proprio quella di trovare almeno un punto di A.

Soprattutto però spesso X ha una struttura geometrica meno restrittiva che permette talvolta una formulazione geometrica degli algoritmi o una riformulazione analitica del problema.

Se l'insieme A non è finito, l'esistenza del minimo non è ovvia; è garantita però, come è noto, se A è un sottoinsieme compatto di $\mathbb{R}^n$ e la funzione f è continua.

Negli ultimi 20 anni la teoria dell'ottimizzazione è diventato un campo molto difficile della matematica, con tecniche prese dalla teoria dei grafi, dalla complicatissima geometria convessa, dalla topologia in 4 dimensioni, dalla geometria dei numeri e dalla programmazione logica con vincoli.

W. Alt: Nichtlineare Optimierung. Vieweg 2002.

M. Grötschel/L. Lovász/A. Schrijver: Geometric algorithms and combinatorial optimization. Springer 1988.

C. Großmann/J. Terno: Numerik der Optimierung. Teubner 1997.

P. Gruber/J. Wills (ed.): Handbook of convex geometry. 2 volumi. North-Holland 1993.

F. Jarre/J. Stoer: Optimierung. Springer 2004.

A. Joereßen/H. Sebastian: Problemlösung mit Modellen und Algorithmen. Teubner 1998.

D. Jungnickel: Optimierungsmethoden. Springer 1999.

K. Marriott/P. Stuckey: Programming with constraints. MIT Press 1998.

K. Neumann/M. Morlock: Operations research. Hanser 1993.

A. Schrijver: Theory of linear and integer programming. Wiley 1990.

Ottimizzazione genetica

Gli algoritmi genetici sono una famiglia di tecniche di ottimizzazione che si ispirano all'evoluzione naturale. I sistemi biologici sono il risultato di processi evolutivi basati sulla riproduzione selettiva degli individui migliori di una popolazione sottoposta a mutazioni e ricombinazione genetica. L'ambiente svolge un ruolo determinante nella selezione naturale in quanto solo gli individui più adatti tendono a riprodursi, mentre quelli le cui caratteristiche sono meno compatibili con l'ambiente tendono a scomparire.

L'ottimizzazione genetica può essere applicata a problemi le cui soluzioni sono descrivibili mediante parametri codificabili capaci di rappresentarne le caratteristiche essenziali. Il ruolo dell'ambiente viene assunto dalla funzione obiettivo che deve essere ottimizzata.

Questo metodo presenta due grandi vantaggi: non dipende da particolari proprietà matematiche e soprattutto la complessità è in generale praticamente lineare. Negli algoritmi genetici, dopo la generazione iniziale di un insieme di possibili soluzioni (individui), alcuni individui sono sottoposti a mutazioni e a scambi di materiale genetico. La funzione di valutazione determina quali dei nuovi individui possono sostituire quelli originali.

Questa tecnica viene applicata con successo a problemi di ricerca operativa, al raggruppamento automatico (un campo della statistica che si occupa di problemi di raggruppamento e classificazione di dati), al problema del commesso viaggiatore, all'approssimazione di serie temporali, alla previsione della conformazione spaziale di proteine a partire dalla sequenza degli aminoacidi, all'ottimizzazione di reti neuronali e di sistemi di Lindenmayer, a modelli di vita artificiale (sociologi tentano invece di simulare l'evoluzione di comportamenti, ad esempio tra gruppi sociali o nazioni).

Nell'applicazione di questi metodi il matematico può intervenire in vari modi: nello sviluppo e nel controllo degli algoritmi (generazione di numeri casuali per la ricerca di conformazioni ottimali in uno spazio multidimensionale di conformazioni, grafica al calcolatore), nella codifica dei dati, nell'organizzazione delle informazioni.

Un campo di ricerca piuttosto attivo è l'ottimizzazione genetica di programmi al calcolatore (il linguaggio più adatto è, per la semplicità della sua sintassi fondamentale, il Lisp), una tecnica che viene detta *programmazione genetica* (in inglese *genetic programming*) e rientra nell'ambito dell'apprendimento di macchine (in inglese *machine learning*).

L'algoritmo di base

Come vedremo, nell'ottimizzazione genetica è molto importante studiare bene la struttura interna del problema e adattare l'algoritmo utilizzato alle caratteristiche del problema. Nonostante ciò presentiamo qui un algoritmo di base che può essere utilizzato in un primo momento e che ci servirà anche per le applicazioni al raggruppamento automatico.

Siano dati un insieme X e una funzione $f : X \longrightarrow \mathbb{R}$. Vogliamo minimizzare f su X (nell'ottimizzazione genetica i vincoli devono in genere essere descritti dalla funzione f stessa e quindi l'insieme ammissibile A coincide con X).

Fissiamo una grandezza n della popolazione, non troppo grande, ad esempio un numero tra 40 e 100. L'algoritmo consiste dei seguenti passi:

(1) Viene generata in modo casuale una popolazione P di n elementi di X.

(2) Per ciascun elemento x di P viene calcolato il valore $f(x)$ (detto *rendimento* di x).

(3) Gli elementi di P vengono ordinati in ordine crescente secondo il rendimento (in ordine crescente perché vogliamo minimizzare il rendimento, quindi gli elementi migliori sono quelli con rendimento minore).

(4) Gli elementi migliori vengono visualizzati sullo schermo oppure il programma controlla automaticamente se i valori raggiunti sono soddisfacenti.
In questo punto l'algoritmo può essere interrotto dall'osservatore o dal programma.

(5) Gli elementi peggiori (ad esempio gli ultimi 10) vengono sostituiti da nuovi elementi generati in modo casuale.

(6) Incroci.

(7) Mutazioni.

(8) Si torna al punto 2.

Gli algoritmi genetici si basano quindi su tre operazioni fondamentali: *rinnovamento* (introduzione di nuovi elementi nella popolazione), *mutazione*, *incroci*.

Confronto con i metodi classici

Il processo evolutivo è un processo lento, quindi se la funzione da ottimizzare è molto regolare (differenziabile o convessa), gli algoritmi classici aprossimano la soluzione molto più rapidamente e permettono una stima dell'errore. Ma in molti problemi pratici, in cui la funzione di valutazione è irregolare o complicata (se ad esempio dipende in modo non lineare da molti parametri) e non accessibile ai metodi tradizionali, l'ottimizzazione genetica può essere di grande aiuto.

Sul significato degli incroci

Le mutazioni da sole non costituiscono un vero *algoritmo*, ma devono essere considerate come un più o meno abile meccanismo di ricerca casuale. Naturalmente è importante lo stesso che anche le mutazioni vengano definite nel modo più appropriato possibile.

Sono però gli incroci che contribuiscono la caratteristica di algoritmo, essenzialmente attraverso un meccanismo di *divide et impera*. Per definirle nel modo più adatto bisogna studiare attentamente il problema, cercando di individuarne *componenti* che possono essere variati indipendentemente l'uno dagli altri, cioè in modo che migliorando il rendimento di un componente non venga diminuito il rendimento complessivo.

Ciò non è sempre facile e richiede una buona comprensione del problema per arrivare possibilmente a una sua riformulazione analitica. Il matematico può contribuire notevolmente in questa analisi e nella formulazione del modello.

Il metodo spartano

Il criterio di scelta adottato dalla selezione naturale predilige in ogni caso gli individui migliori, dando solo ad essi la possibilità di moltiplicarsi. Questo meccanismo tende a produrre una certa uniformità qualitativa in cui i progressi possibili diventano sempre minori e meno probabili. Il risultato finale sarà spesso una situazione apparentemente ottimale e favorevole, ma incapace di consentire altri miglioramenti, un *ottimo locale*.

Perciò non è conveniente procedere selezionando e moltiplicando in ogni passo solo gli elementi migliori, agendo esclusivamente su di essi con mutazioni e incroci. Se si fa così infatti dopo breve tempo le soluzioni migliori risultano tutte imparentate tra loro ed è molto alto il rischio che l'evoluzione stagni in un ottimo locale che interrompe il processo di adattamento senza consentire ulteriori miglioramenti essenziali.

Per questa ragione, per impedire il proliferare di soluzioni tutte imparentate tra di loro, a differenza dalla selezione naturale non permettiamo la proliferazione identica. Nelle *mutazioni* il peggiore tra l'originale e il mutante viene sempre eliminato, e negli *incroci* i due nuovi elementi sostituiscono entrambi i vecchi, anche se solo uno dei due nuovi è migliore dei vecchi.

Precisiamo quest'ultimo punto. Supponiamo di voler incrociare due individui A e B della popolazione, rappresentati come coppie di componenti che possono essere scambiati: $A = (a_1, a_2), B = (b_1, b_2)$. Gli incroci ottenuti siano per esempio $A' = (a_1, b_2), B' = (b_1, a_2)$. Calcoliamo i rendimenti e assumiamo che i migliori due dei quattro elementi siano A' e B. Se però scegliamo questi due, nelle componenti abbiamo (a_1, b_2) e (b_1, b_2) e vediamo che il vecchio B è presente in 3 componenti su 4 e ciò comporterebbe quella propagazione di parentele che vogliamo evitare.

Negli incroci seguiamo quindi il seguente principio: Se nessuno dei due nuovi elementi è migliore di entrambi gli elementi vecchi, manteniamo i vecchi e scartiamo gli incroci; altrimenti scartiamo entrambi gli elementi vecchi e manteniamo solo gli incroci.

Numeri casuali

Successioni di numeri (o vettori) casuali (anche in forme di tabelle) vengono usate da molto tempo in problemi di simulazione, statistica, integrazione numerica e crittografia. Attualmente esiste un grande bisogno di tecniche affidabili per la generazione di numeri casuali, come mostra l'intensa ricerca in questo campo che impiega spesso tecniche complicate della teoria dei numeri.

Il termine numero casuale ha tre significati. Esso, nel calcolo delle probabilità, denota una *variabile casuale* a valori numerici (reali o interi), cioè un'entità che non è un numero ma, nell'assiomatica di Kolmogorov, una funzione misurabile nel senso di Borel a valori reali (o a valori in $\mathbb{R}^n$ quando si tratta di vettori casuali) definita su uno spazio di probabilità, mentre le successioni generate da metodi matematici, le quali sono per la loro natura non casuali ma deterministiche, vengono tecnicamente denominate successioni di numeri *pseudocasuali*.

Il terzo significato è quello del linguaggio comune, che può essere applicato a numeri ottenuti con metodi *analogici* (dadi, dispositivi meccanici o elettronici ecc.), la cui casualità però non è sempre affidabile (ad esempio per quanto riguarda il comportamento a lungo termine) e le cui proprietà statistiche sono spesso non facilmente descrivibili (di un dado forse ci possiamo fidare, ma un dispositivo più complesso può essere difficile da giudicare). Soprattutto per applicazioni veramente importanti è spesso necessario creare una quantità molto grande di numeri casuali, e a questo scopo non sono sufficienti i metodi analogici. Oltre a ciò normalmente bisogna conoscere a priori le proprietà statistiche delle successioni che si utilizzano.

Siccome solo le successioni ottenute con un algoritmo deterministico si prestano ad analisi di tipo teorico, useremo spesso il termine „numero casuale“ come abbreviazione di „numero pseudocasuale“.

Una differenza importante anche nelle applicazioni è che per le successioni veramente casuali sono possibili soltanto stime probabilistiche, mentre per le successioni di numeri pseudocasuali si possono ottenere, anche se usualmente con grandi difficoltà matematiche, delle stime precise.

Spieghiamo l'importanza di questo fatto assumendo che il comportamento di un dispositivo importante (che ad esempio governi un treno o un missile) dipenda dal calcolo di un complicato integrale multidimensionale che si è costretti ad eseguire mediante un metodo di Monte Carlo.

Se i numeri casuali utilizzati sono analogici, cioè veramente casuali, allora si possono dare soltanto stime per la probabilità che l'errore non superi una certa quantità permessa, ad esempio si può soltanto arrivare a poter dire che in non più di 15 casi su 100000 l'errore del calcolo sia tale da compromettere le funzioni del dispositivo. Con successioni pseudocasuali (cioè generate da metodi matematici), le stime di errore valgono invece in tutti i casi, e quindi si può garantire che l'errore nel calcolo dell'integrale sia sempre minore di una quantità fissa, assicurando così che il funzionamento del dispositivo non venga mai compromesso.

---

runif

Una successione di n numeri casuali reali (uniformemente distribuiti) in $[a, b]$ si ottiene con

```
runif(n,min=a,max=b)
```

Si possono ottenere anche numeri casuali distribuiti secondo una distribuzione normale con `rnorm`. Nell'ottimizzazione genetica spesso vogliamo anche usare numeri casuali interi; a questo scopo definiamo la seguente funzione:

```
Snc.interi = function (n,min=1,max=6)
floor(runif(n,min=min,max=max+0.999))
```

R usa, nell'impostazione iniziale, come generatore di numeri casuali un algoritmo detto *Mersenne twister*, dovuto a Matsumoto e Nishimura, considerato uno dei migliori generatori conosciuti.

Numeri casuali in crittografia

Si dice che Cesare abbia talvolta trasmesso messaggi segreti in forma crittata, facendo sostituire ogni lettera dalla terza lettera successiva (quindi la a dalla d, la b dalla e, ..., la z dalla c), cosicché *crascastramovebo* diventava *fudvfdvwudpryher* (usando il nostro alfabeto di 26 lettere). È chiaro che un tale codice è facile da decifrare. Se invece $(x_1, \ldots, x_N)$ è una successione casuale di interi tra 0 e 25 e il testo $a_1a_2...a_N$ viene sostituito da $a_1{+}x_1,...,a_N{+}x_N$, questo è un metodo sicuro. Naturalmente sia il mittente che il destinatario devono essere in possesso della stessa lista di numeri casuali.

La scoperta dei farmaci

È poco noto che il numero dei bersagli molecolari dei farmaci attualmente prodotti è piuttosto piccolo (circa 500) e che lo sviluppo di una nuova sostanza farmaceutica consuma somme ingenti (400 milioni di euro per una nuova molecola). Oltre ai bersagli classici (recettori sulle membrane cellulari, enzimi e recettori ormonali) in futuro avranno sempre più importanza i bersagli legati al genoma e ciò implicherà, secondo le previsioni, un probabile aumento dei bersagli a molte migliaia in pochi anni.

È forse sorprendente che in un recente testo di disegno dei farmaci si trovi il seguente brano che abbiamo tradotto: *„I matematici negli ultimi decenni hanno aggiunto i principi dell'evoluzione al loro strumentario. Utilizzando replicazioni, mutazioni e incroci essi hanno sviluppato algoritmi genetici. Chi ha mai potuto ammirare come un tale algoritmo risolve i più complessi problemi di ottimizzazione in tempo incredibilmente breve, non avrà più dubbi che anche l'evoluzione delle specie biologiche si è svolta in modo analogo.“* (Böhm/Klebe/Kubinyi, 231)

Il futuro dell'industria farmaceutica sarà fortemente influenzato dai progressi nella comprensione dettagliata delle informazioni contenute nel genoma e della struttura e funzione delle molecole biologiche per i processi normali e patologici della vita, e quindi anche una sempre migliore comprensione molecolare delle malattie che permetterà una progettazione razionale e mirata di molecole farmaceutiche. Nuove tecniche permettono di fornire in tempi brevi numerosi composti da sottoporre a test e da classificare; il matematico, nel suo ruolo di semplificatore della complessità, può nella ricerca sviluppare nuovi metodi di classificazione o nuovi test statistici.

H. Böhm/G. Klebe/H. Kubinyi: Wirkstoffdesign. Spektrum 1996.
M. Vose: The simple genetic algorithm. MIT Press 1999.

IX. RAGGRUPPAMENTO AUTOMATICO

Analisi di gruppi

In un campione di dati statistici sono spesso presenti gruppi che possono essere noti in anticipo o meno. Dell'analisi di queste strutture si occupano soprattutto tre grandi discipline statistiche: l'analisi della varianza, l'analisi delle discriminanti e la teoria del raggruppamento automatico.

Nell'*analisi della varianza* la suddivisione in gruppi è già nota e si studia se e come una o più variabili statistiche differiscano da un gruppo all'altro. Nel caso di una variabile si parla di analisi della varianza univariata, nel caso di più variabili di analisi della varianza multivariata.

Anche nell'*analisi delle discriminanti* la suddivisione in gruppi è nota e si cercano funzioni discriminanti con cui distinguere i gruppi. Assumiamo quindi di avere un insieme di pazienti $A \subset \mathbb{R}_m$ e una partizione $A = S \sqcup M$ in sani e malati. Allora cerchiamo una funzione $f : \mathbb{R}_m \longrightarrow R$, detta *funzione discriminante*, tale che gli insiemi degli individui con test positivo risp. negativo corrispondano il più possibile ad M ed S. Spesso l'insieme dei positivi è definito come $P := (f > 0)$ e quindi l'insieme dei negativi come $N := (f \leq 0)$. È importante che nelle applicazioni dell'analisi delle discriminanti in statistica medica in genere si vorrebbe successivamente applicare lo stesso criterio f a individui che non fanno parte di A per poter valutare se siano affetti da quella malattia.

Nella terza delle tre discipline, la teoria del *raggruppamento automatico* (nella letteratura inglese nota come *cluster analysis*), non è ancora nota la suddivisione in gruppi e l'obiettivo è proprio un tale raggruppamento. Ci occuperemo di questo compito in questa parte finale del corso.

Raggruppamento automatico

Questo campo della statistica si occupa della costruzione di raggruppamenti (in inglese *cluster* significa grappolo, gruppetto) da un insieme di dati ed è particolarmente adatto per l'uso degli algoritmi genetici, sia perché mutazioni e incroci sono definibili in modo molto naturale, sia perché nella cluster analysis viene utilizzata una molteplicità di criteri di ottimalità per le partizioni che negli approcci tradizionali richiedono ogni volta algoritmi di ottimizzazione diversi e spesso computazionalmente difficili e quindi non applicabili per insiemi grandi (e spesso anche solo medi) di dati, mentre, come abbiamo già osservato, gli algoritmi genetici non dipendono dalle proprietà matematiche delle funzioni utilizzate e hanno una complessità che cresce solo in modo lineare con il numero dei dati. Siccome l'algoritmo non dipende dalla funzione di ottimalità scelta, anche se ci limiteremo probabilmente all'uso del cosiddetto criterio della varianza, lo stesso algoritmo può essere usato per un criterio di ottimalità qualsiasi. Nella letteratura è descritta una grande varietà di misure di somiglianza o di diversità, tra le quali in un'applicazione concreta si può scegliere per definire l'ottimalità delle partizioni, ma il modo in cui viene usato l'algoritmo genetico è sempre uguale.

È per esempio piuttosto difficile trovare algoritmi tradizionali per il caso che l'omogeneità e la diversità dei gruppi non siano descritte mediante misure di somiglianza o diversità tra gli individui ma direttamente da misure per i gruppi, mentre ciò non causa problemi per l'algoritmo genetico.

Elenchiamo alcuni campi di applicazione del raggruppamento automatico: classificazione di specie in botanica e zoologia (*tassonomia numerica*) o di aree agricole o biogeografiche, classificazione di specie virali o batteriche, definizione di gruppi di persone con comportamento (istruzione, attitudini, ambizioni, livello di vita, professione) simile in studi sociologici o psicologici, creazione di gruppi di dati omogenei nell'elaborazione dei dati (per banche dati o grandi biblioteche), elaborazione di immagini (ad esempio messa in evidenza di formazioni patologiche in radiografie mediche), individuazione di gruppi di pazienti con forme diverse di una malattia o riguardo alla risposta a un tipo di trattamento, classificazione di malattie in base a sintomi e test di laboratorio, studi linguistici, raggruppamento di regioni (province, comuni) relativamente a caratteristiche economiche (o livello generale di vita o qualità dei servizi sanitari), individuazione di gruppi di località con frequenza simile per quanto riguarda una determinata malattia, reperti archeologici o paleontologici o mineralogici (descritti ad esempio mediante la loro composizione chimica) o antropologici, dati criminalistici (impronte digitali, caratteristiche genetiche, forme di criminalità e loro distribuzione geografica o temporale), confronto tra molecole organiche, classificazione di scuole pittoriche, indagini di mercato (in cui si cerca di individuare gruppi omogenei di consumenti), raggruppamenti dei clienti di un assicurazione in gruppi per definire il prezzo delle polizze, classificazione di strumenti di lavoro o di prodotti nell'industria oppure dei posti di lavoro in una grande azienda, confronto del costo della vita nei paesi europei, divisione dei componenti di un computer in gruppi per poterli disporre in modo da minimizzare la lunghezza di cavi e circuiti.

In queste applicazioni, che si differenziano fortemente per la quantità degli oggetti da classificare (poche decine nel caso di oggetti archeologici, milioni di pixel nell'elaborazione di immagini) e per la natura dei dati, spesso non è facile scegliere un criterio di ottimalità robusto (cambi di scala possono ad esempio influenzare l'esito della classificazione, quando si usano distanze euclidee) e superare la spesso notevole complessità computazionale.

Il criterio della varianza

A sia un sottoinsieme finito di $\mathbb{R}_m$.
Per un sottoinsieme non vuoto α di A denotiamo con

$$\overline{\alpha} := \frac{1}{|\alpha|} \sum_{x \in \alpha} x$$

il baricentro di α. Poniamo inoltre

$$\Delta\alpha := \sum_{x \in \alpha} |x - \overline{\alpha}|^2$$

Per una partizione P di A sia infine

$$g(P) := \sum_{\alpha \in P} \Delta\alpha$$

Questa è la funzione da minimizzare quando si usa *il criterio della varianza*.

Più precisamente si fissa il numero k delle classi della partizione; la partizione ottimale è quella partizione P di A con k classi per cui $g(P)$ assume il minimo; il minimo esiste certamente, perché A è un insieme finito e quindi anche il numero delle partizioni di A è finito, benché molto grande.

In generale, nel raggruppamento automatico si vorrebbe da un lato che ogni classe della partizione sia il più possibile omogenea e quindi le distanze tra gli elementi di una stessa classe siano piccole, dall'altro che le classi siano il più separate tra di loro. Il criterio della varianza soddisfa, come si può dimostrare, allo stesso tempo entrambe queste richieste. Esso è, per dati che hanno una rappresentazione naturale nel $\mathbb{R}_m$, il criterio di ottimalità più usato, benché non esente da limitazioni (cfr. pagina 40); bisogna in ogni caso come sempre scalare in modo appropriato le variabili, utilizzando ad esempio una delle tecniche di standardizzazione che conosciamo.

Suddivisione gerarchica

La teoria dei raggruppameni comprende numerose tecniche e oltre a raggruppamenti tramite partizioni si utilizzano anche *ricoprimenti* (cioè rappresentazioni dell'insieme dei dati come unione di insiemi non necessariamente disgiunti) e *suddivisioni gerarchiche* (spesso rappresentate tramite *dendrogrammi*). Queste ultime sono usate frequentemente nella letteratura statistica applicata, ma spesso in modo non appropriato; è infatti difficile la loro corretta interpretazione. La teoria matematica della classificazione gerarchica si basa sulle *metriche non archimedee* (o *ultrametriche*) ed è esposta nei libri di Diday/ e Bock. Ultrametriche sono note e utilizzate da molto tempo in matematica, soprattutto in alcuni campi dell'algebra e della teoria dei numeri e nella dinamica simbolica.

Una metrica d si dice non archimedea, se per ogni numero reale $\varepsilon > 0$ vale la relazione di transitività

$$d(a,b) < \varepsilon \text{ e } d(b,c) < \varepsilon \Longrightarrow d(a,c) < \varepsilon$$

Ciò significa che la relazione

$$a \sim_\varepsilon b \iff d(a,b) < \varepsilon$$

(riflessiva e simmetrica per ogni metrica) è una relazione di equivalenza. Questa condizione, molto naturale nella statistica, non è soddisfatta nella geometria euclidea: se la distanza tra a e b è minore di un metro e lo stesso vale per la distanza tra b e c, da ciò non segue che anche la distanza tra a e c sia minore di un metro. Metriche non archimedee non misurano una distanza geometrica, ma comunanze: più proprietà due oggetti hanno in comune, più simili e vicini risultano in un'appropriata metrica non archimedea.

Il numero delle partizioni

Quante sono le partizioni di un insieme finito? Denotiamo con $S(n,k)$ il numero delle partizioni di un insieme con n elementi in k classi. I numeri della forma $S(n,k)$ sono detti *numeri di Stirling di seconda specie*.

Lemma 38.1. *Per $n,k \geq 1$ vale*

$$S(n,k) = S(n-1,k-1) + k \cdot S(n-1,k)$$

Dimostrazione. Una partizione di $\{1,\dots,n\}$ può contenere $\{n\}$ come elemento (in tal caso n è equivalente solo a se stesso) oppure no.

Il numero delle partizioni di $\{1,\dots,n\}$ in k classi di cui una coincide con $\{n\}$ è evidentemente uguale al numero delle partizioni di $\{1,\dots,n-1\}$ in $k-1$ classi, cioè uguale a $S(n-1,k-1)$.

Se una partizione di $\{1,\dots,n\}$ con k classi non contiene $\{n\}$ come elemento, essa si ottiene da una partizione di $\{1,\dots,n-1\}$ in k classi, aggiungendo n ad una delle k classi. Per fare questo abbiamo k possibilità.

Dalla definizione otteniamo direttamente le seguenti relazioni (per la prima si osservi che l'insieme vuoto $\emptyset$ può essere considerato in modo banale come partizione di $\emptyset$).

$$S(0,0) = 1.$$
$$S(0,k) = 0 \text{ per } k \geq 1.$$
$$S(n,0) = 0 \text{ per } n \geq 1$$

Possiamo così scrivere un programma in R per il calcolo ricorsivo di $S(n,k)$:

```
M.stirling2 = function (n,k)
{if (n==0) if (k==0) 1 else 0
else if (k==0) 0 else
Recall(n-1,k-1)+k*Recall(n-1,k)}
```

I numeri di Stirling di seconda specie crescono fortemente:

$$S(5,2) = 15$$
$$S(10,2) = 511$$
$$S(10,3) = 9330$$
$$S(20,5) = 749206090500$$
$$S(50,4) = 52818655359845226611906445312$$

Calcolo della funzione g

Rappresentiamo in primo luogo il sottoinsieme A mediante la matrice dei dati X in $\mathbb{R}^n_m$; più precisamente A è l'insieme delle righe di X. Denotiamo con k il numero delle classi. Una partizione è rappresentata da un vettore $P \in \{1,\dots,k\}^n$. Una riga X^i appartiene alla a-esima classe α_a se P^i è uguale ad a.

Per ogni $a \in \{1,\dots,k\}$ dobbiamo calcolare il baricentro $\overline{\alpha_a}$; otteniamo così una matrice $B \in \mathbb{R}^k_m$ con $B^a = \overline{\alpha_a}$, almeno se la a-esima classe non è vuota, perché altrimenti il baricentro $\overline{\alpha_a}$ non è ben definito. D'altra parte però gli indici a con $\alpha_a = \emptyset$ non entrano veramente nel calcolo di g, come risulta da

$$g(P) = \sum_{\alpha \in P} \sum_{x \in \alpha} |x - \overline{\alpha}|^2$$

o dalla formula equivalente

$$g(P) = \sum_{i=1}^{n} |X^i - B^{P^i}|^2$$

Infatti, se $\alpha_a = \emptyset$, P^i sarà sempre $\neq a$. Qui possiamo utilizzare a nostro favore il fatto che R permette di creare matrici numeriche in cui appaiono i valori `Inf` e `NaN`, per cui possiamo creare una matrice B che contiene anche questi valori come elementi.

Definiamo prima una funzione che per ogni vettore $v \in \{1,\dots,k\}^n$ calcola la frequenza con cui appaiono i suoi elementi:

```
S.conta = function (v,k)
{u=rep(0,k)
for (a in v) u[a]=u[a]+1
u}

# Esempio:
v=c(1,2,4,4,1,1,4,2,5)
print(v)
# 1 2 4 4 1 1 4 2 5

u=S.conta(v,5)
print(u)
# 3 2 0 3 1
```

Adesso calcoliamo la matrice dei baricentri. Nella penultima riga appare l'espressione `B[a,]/cont[a]` che in R però è ammissibile anche quando il denominatore si annulla. Infatti, quando `cont[a]` è uguale a zero, anche `B[a,]` è uguale a zero, e `0/0` in R diventa `NaN`, valore che, come abbiamo detto, può far parte dei coefficienti di una matrice.

```
Sra.baricentri = function (X,P,k)
{m=ncol(X); n=nrow(X)
cont=S.conta(P,k)
B=Mm(rep(0,m*k),righe=k)
for (i in 1:n)
{a=P[i]; B[a,]=B[a,]+X[i,]}
for (a in 1:k) B[a,]=B[a,]/cont[a]
B}
```

A questo punto possiamo definire la funzione per il calcolo di g:

```
Sra.g = function (X,P,k)
{n=nrow(X); B=Sra.baricentri(X,P,k)
s=0
for (i in 1:n) {u=X[i,]-B[P[i],]
s=s+Mv.scalare(u,u)} s}
```

Consideriamo la prima figura a pagina 32. Potremmo pensare a due partizioni P e Q a tre classi α, β e γ.

Nella partizione P poniamo

α = {To, Mi, Fi, Bo, Pr, Pd}
β = {Ge, Fe, Ve, Ra, Pi}
γ = {Bz,Tn,Bl,Vi}

nella partizione Q spostiamo Pisa da β a γ.

Tenendo conto dell'ordine in cui i comuni appaiono nella tabella a pagina 26, P e Q diventano vettori definiti nella tabella seguente:

	P	Q
Belluno	3	3
Bologna	1	1
Bolzano	3	3
Ferrara	2	2
Firenze	1	1
Genova	2	2
Milano	1	1
Padova	1	1
Parma	1	1
Pisa	2	3
Ravenna	2	2
Torino	1	1
Trento	3	3
Venezia	2	2
Vicenza	3	3

Come standardizzazione usiamo di nuovo la matrice dei ranghi. Quale delle due partizioni è migliore?

```
Db(2)
X=Db.matrice()
XR=Sm.rango(X)

P=c(3,1,3,2,1,2,1,1,1,2,2,1,3,2,3)
Q=c(3,1,3,2,1,2,1,1,1,3,2,1,3,2,3)

gp=Sra.g(P,XR,3)
gq=Sra.g(Q,XR,3)

print(gp)
# 2.141156

print(gq)
# 2.63733
```

La partizione P è quindi migliore. Con la matrice non standardizzata invece risulterebbe leggermente migliore la seconda partizione:

```
gp=Sra.g(P,X,3)
gq=Sra.g(Q,X,3)

print(gp)
# 1484101

print(gq)
# 1448306
```

Il programma principale

Presentiamo adesso un programma completo in R che contiene le funzioni per il raggruppamento automatico mediante un algoritmo genetico. Il programma è piuttosto semplice e segue l'algoritmo di base dell'ottimizzazione genetica visto a pagina 35. Benché molto più lento di un programma analogo in C, è sufficiente per trattare i nostri 15 comuni.

La funzione principale `Sra` è interattiva, permettendo all'utente di impostare durante l'esecuzione l'intervallo di tempo `dt` che intercorre tra le visualizzazioni del risultato ottimale raggiunto.

```
Sra = function (X,k)
{n=nrow(X)
MP=Mm(numeric(n*40),col=40)
MP=Sra.nuovi(MP,k,1,40)
dt=10; t=0; repeat
{t=t+1; R=Sra.rendimenti(MP,X,k)
Ord=order(R); MP=MP[,Ord]
if (t%%dt==0)
{dt=Sra.visualizza(t,dt,
  R[Ord[1]],MP[,1])
if (dt==0) break}
MP=Sra.nuovi(MP,k,31,40)
MP=Sra.mutazioni(MP,k,R,X)
MP=Sra.incroci(MP,k,R,X)}}
```

Si noti l'introduzione del vettore `R` dei rendimenti. Per le visualizzazioni usiamo

```
Sra.visualizza = function (t,dt,rend,P)
{P=paste(P,collapse=' ')
cat('\n',rend,': ',P,'   dopo ',t,
  ' generazioni\n',sep='')
a=readline('Vuoi continuare? ')
if (a=='n') 0
else {v=as(a,'numeric')
if (!is.na(v)) v else dt}}
```

Battendo semplicemente invio, il programma continua; con `'n'` si ferma, mentre se inseriamo un numero, questo viene usato come nuovo valore della variabile `dt` che indica l'intervallo tra due visualizzazioni.

L'algoritmo genetico

La creazione di una nuova *matrice di partizioni* (MP), che contiene 40 colonne ciascuna delle quali rappresenta una partizione, avviene con

```
Sra.nuovi = function (MP,k,a,b)
{n=nrow(MP); for (j in a:b)
MP[,j]=Snc.interi(n,1,k)
MP}
```

il calcolo dei rendimenti con

```
Sra.rendimenti = function (MP,X,k)
apply(MP,2,Sra.g,X,k)
```

Qui viene usata la funzione `Sra.g` definita a pagina 38.

Per le mutazioni usiamo

```
Sra.muta = function (P,k,R)
{n=length(P); p=runif(1,0,0.5)
for (i in 1:n) if (runif(1)<p)
P[i]=Snc.interi(1,1,k)
P}
```

e

```
Sra.mutazioni = function (MP,k,R,X)
{for (j in 1:40)
{P=Sra.muta(MP[,j],k,R)
if (Sra.g(P,X,k)<R[j]) MP[,j]=P}
MP}
```

per gli incroci

```
Sra.incroci = function (MP,k,R,X)
{n=nrow(MP)
for (j in seq(1,40,2))
{P1=MP[,j]; P2=MP[,j+1]
p=runif(1,0,0.5)
for (i in 1:n) if (runif(1)<p)
{a=P1[i]; P1[i]=P2[i]; P2[i]=a}
R1=Sra.g(P1,X,k); R2=Sra.g(P2,X,k)
if ((R1<R[j])||(R2<R[j+1]))
{MP[,j]=P1; MP[,j+1]=P2}}
MP}
```

Vengono incrociate la prima con la seconda partizione, la terza con la quarta, e così via. Nelle mutazioni e negli incroci applichiamo il metodo spartano.

Raggruppamenti dei 15 comuni

Applichiamo il metodo ai 15 comuni. Chiediamo un raggruppamento in 4 classi ed eseguiamo l'algoritmo prima senza standardizzazione con

```
Db(2)
X=Db.matrice()
Sra(X,4)
```

usando, con il comando `Db(2)`, la nostra banca dati. Dopo 200 generazioni otteniamo il risultato

```
410593.9: 1 2 1 3 2 2 4 1 3 1 3 4 1 3 1
```

Proviamo la proiezione su $[0, 1]$:

```
Db(2)
X=Db.matrice()
X01=Sm.tra01(X)
Sra(X01,4)
```

Dopo 200 generazioni otteniamo

```
1.124847: 1 2 1 3 2 2 4 2 2 2 3 4 1 3 2
```

Si noti che i rendimenti non sono confrontabili (perché abbiamo usato standardizzazioni diverse) e possono essere usati solo per valutare la bontà del risultato per esecuzioni con la stessa standardizzazione.

Nello stesso modo procediamo per la matrice dei ranghi:

```
Db(2)
X=Db.matrice()
XR=Sm.rango(X)
Sra(XR,4)
```

ottenendo dopo 200 generazioni

```
1.565391: 3 4 3 1 4 1 4 2 4 2 1 4 3 1 2
```

Per vedere concretamente le partizioni riportiamo i risultati in una tabella:

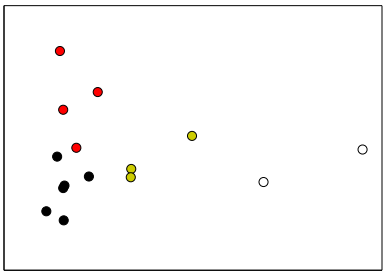
	X	X01	XR
Belluno	1	1	3
Bologna	2	2	4
Bolzano	1	1	3
Ferrara	3	3	1
Firenze	2	2	4
Genova	2	2	1
Milano	4	4	4
Padova	1	2	2
Parma	3	2	4
Pisa	1	2	2
Ravenna	3	3	1
Torino	4	4	4
Trento	1	1	3
Venezia	3	3	1
Vicenza	1	2	2

Si osservi che i numeri delle partizioni possono essere permutati tra di loro e che perciò il gruppo 1 e il gruppo 2 non sono più simili di quanto lo siano il gruppo 1 e il gruppo 4. Abbiamo così i seguenti gruppi.

Senza standardizzazione:

Belluno, Bolzano, Padova, Pisa, Trento, Vicenza
Bologna, Firenze, Genova
Ferrara, Parma, Ravenna, Venezia
Milano, Torino

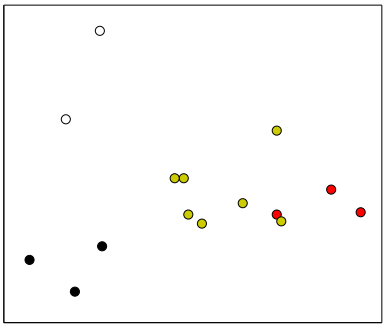
Usando le figure alle pagine 30-32, coloriamo i comuni in colori diversi a seconda della classe nella partizione generata dall'algoritmo di raggruppamento.



Quando confrontiamo i risultati, dobbiamo ricordarci che si tratta di proiezioni 2-dimensionali, mentre il raggruppamento avviene (nel nostro caso) in quattro dimensioni. Questo spiega perché ad esempio nella prossima figura un punto giallo è apparentemente (cioè in due dimensioni) separato dagli altri punti della stessa classe.

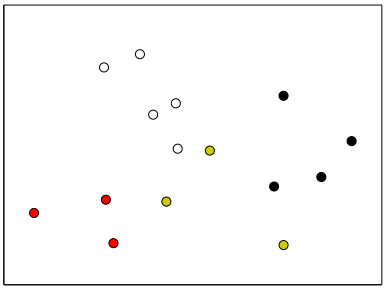
Con proiezione su $[0, 1]$:

Belluno, Bolzano, Trento
Bologna, Firenze, Genova, Padova, Parma, Pisa, Vicenza
Ferrara, Ravenna, Venezia
Milano, Torino



Con la matrice dei ranghi:

Ferrara, Genova, Ravenna, Venezia
Padova, Pisa, Vicenza
Belluno, Bolzano, Trento
Bologna, Firenze, Milano, Parma, Torino



Soprattutto in problemi complicati i risultati di un'ottimizzazione genetica non sono unici e possono differire da un'esecuzione all'altra, anche dopo lo stesso numero di generazioni.

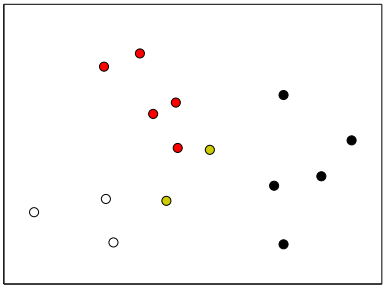
Nell'ultimo esempio (in cui si usa la matrice dei ranghi) può sorprendere che Genova si trovi nello stesso gruppo di Ferrara; perciò proviamo un'altra esecuzione, trovando dopo 400 generazioni

```
1.477381: 4 3 4 1 3 1 3 2 3 1 1 3 4 1 2
```

risultato che rimane uguale anche dopo 800 generazioni e quindi probabilmente è ottimale; esso corrisponde alla partizione

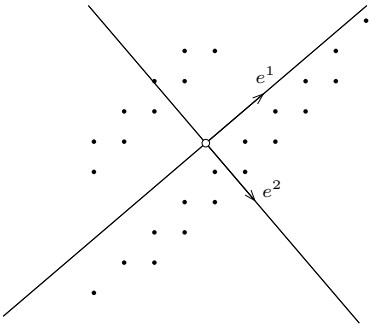
Ferrara, Genova, Pisa, Ravenna, Venezia
Padova, Vicenza
Bologna, Firenze, Milano, Parma, Torino
Belluno, Bolzano, Trento

Genova è rimasta nel gruppo di Ferrara, a cui si è aggiunta Pisa.



Il problema dei gruppi sferici

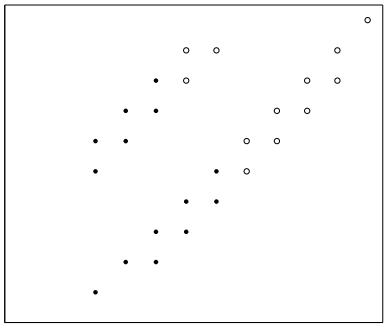
Talvolta un tentativo di raggruppamento automatico parte da dati trasformati mediante un'analisi delle componenti principali. Però non sempre la prima componente principale è la più adatta nei compiti di classificazione. Guardiamo la seguente figura:



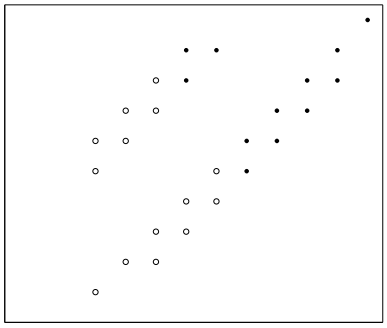
È evidente che la varianza in direzione e^1 è notevolmente maggiore che in direzione e^2; nonostante ciò i dati si distinguono in due gruppi che sono determinati dalla seconda componente principale. Se ciò accade in una proiezione $\mathbb{R}_m \longrightarrow \mathbb{R}_2$ con $m > 2$, una tale divisione in gruppi può sfuggirci.

Ma questo semplice problema di classificazione non si risolve nemmeno con il raggruppamento automatico.

Scopriamo infatti adesso un difetto piuttosto spiacevole dei metodi di raggruppamento o almeno del criterio della varianza, che ne limita in alcune situazioni l'applicazione. Vengono infatti preferiti *gruppi sferici*, anche quando una suddivisione diversa sembrerebbe migliore. Consideriamo l'insieme dei dati nella figura soprastante. Applichiamo il nostro metodo alla matrice non standardizzata con due classi, colorando gli elementi dei due gruppi in modo diverso. Dopo 400 generazioni otteniamo



L'algoritmo ha creato due gruppi approssimativamente sferici invece della più naturale divisione diagonale. Anche una standardizzazione (ad esempio proiezione su $[0, 1]$) ovviamente non elimina il problema:



Abbiamo ottenuto esattamente la stessa partizione!

Bisogna allora provare (se ci si accorge del problema, il che non è sempre facile in dimensione maggiore a 2) ad usare un'altra funzione di ottimalità (ad esempio basata sul *criterio del determinante*), ma anche questa può avere limitazioni a sua volta.

La funzione pam di R

R fornisce il pacchetto `cluster` per le funzioni di raggruppamento automatico. Per ottenere partizioni ottimali si può usare la funzione `pam` che, nella sintassi più semplice, si usa nella forma `pam(tab,k)`, in cui `tab` è una tabella e `k` il numero delle classi.

Dopo `library(cluster)` con

```
Db(2)
tab=Db.tab()
print(pam(tab,4))
```

otteniamo (in un output più complesso) il vettore

```
1 2 1 3 2 2 4 2 3 1 3 4 1 3 1
```

che corrisponde alla partizione

Belluno, Bolzano, Pisa, Trento, Vicenza
Bologna, Firenze, Genova, Padova
Ferrara, Parma, Ravenna, Venezia
Milano, Torino

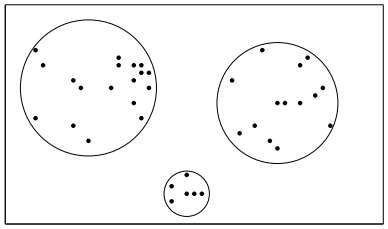
L'esecuzione è molto veloce.

M. Aldenderfer/R. Blashfield: Cluster analysis. Sage 1984.
G. Bahrenberg/E. Giese/J. Nipper: Statistische Methoden in der Geographie II. Borntraeger 2003.
H. Bock: Automatische Klassifikation. Vandenhoeck & Ruprecht 1974.
S. Bolasco: Analisi multidimensionale dei dati. Carocci 2002.
R. Cormack: A review of classification. J. Roy. Stat. Soc. A 134 (1971), 321-367.
P. Diaconis: Group representations in probability and statistics. Hayward 1988.
E. Diday/J. Lemaire/J. Pouget/F. Testu: Eléments d'analyse de données. Dunod 1982.
G. Dunn/B. Everitt: An introduction to mathematical taxonomy. Dover 2004.
H. Eckey/R. Kosfeld/M. Rengers: Multivariate Statistik. Gabler 2002.
L. Fahrmeir/A. Hamerle/G. Tutz: Multivariate statistische Verfahren. De Gruyter 1996.
J. Gentle: Elements of computational statistics. Springer 2002.
J. Hartung/B. Elpelt: Multivariate Statistik. Oldenbourg 1986.
A. Rizzi: Analisi dei dati. NIS 1985.
R. Sokal/P. Sneath: Principles of numerical taxonomy. Freeman 1963.
H. Späth: Clusterformation und -analyse. Oldenbourg 1983.
W. Venables/B. Ripley: Modern applied statistics with S. Springer 2002.

„The major stimulus for the development of clustering methods was a book entitled 'Principles of numerical taxonomy', published in 1963 by two biologists, Robert Sokal and Peter Sneath ... The literature on cluster analysis exploded after the publication of the Sokal and Sneath book ... Despite their popularity, clustering methods are still poorly understood in comparison to such multivariate statistical procedures as factor analysis, discriminant analysis, and multidimensional scaling.“ (Aldenderfer/Blashfield, 7-9)

„In alcuni campi di ricerca si può pertanto ritenere che la fase di classificazione sia il momento essenziale del procedimento scientifico ...“ (Rizzi, 72)

„Do not assume that clustering methods are the best way to discover interesting groupings in the data; in our experience the visualization methods are often far more effective.“ (Venables/Ripley, 316)



Non è sempre così facile.

X. DIFFICOLTÀ IN ALTA DIMENSIONE

I problemi dell'alta dimensione

L'obiettivo della statistica multidimensionale geometrica è di scoprire relazioni tra dati rappresentati da punti in spazi $\mathbb{R}_m$ ad alta dimensione con ad esempio $40 \leq m \leq 100$. Una delle maggiori difficoltà in questo intento è la cosiddetta maledizione dell'alta dimensione (che nella letteratura inglese è nota sotto il termine di *curse of dimensionality*), dovuta soprattutto al fatto che i concetti metrici in spazi a così alte dimensioni perdono gran parte del loro significato perché il volume della palla di raggio 1 in $\mathbb{R}_m$ converge rapidamente a zero; come si vede dalla tabella a destra già nel $\mathbb{R}_{10}$ la palla iscritta occupa solo il 2.5 per mille del volume del cubo. In un cubo ad alta dimensione perciò il volume è concentrato vicino al bordo e ciò crea notevoli problemi per l'interpretazione statistica di considerazioni metriche e gli algoritmi che le utilizzano. Questo capitolo del corso è dedicato alla discussione di queste difficoltà che proprio nella statistica medica, uno dei campi in cui attualmente sono prodotte grandi quantità di dati ad alta dimensione, sono spesso trascurate.

Sfere in $\mathbb{R}_m$

Situazione 1.1. Siano $m \in \mathbb{N}$ ed $r \in \mathbb{R}$ con $r \geq 0$. Indichiamo come finora con m la dimensione dello spazio in cui si trovano i nostri dati. Per $\alpha \in \mathbb{R}$ denotiamo con $[\alpha]$ la parte intera di α.

Definizione 1.2. Per $m > 0$ sia $v_m(r)$ il volume di una palla di raggio r in $\mathbb{R}_m$. Useremo l'abbreviazione

$$v_m := v_m(\tfrac{1}{2})$$

v_m è quindi il volume di una palla *iscritta* a un cubo di lato 1 in $\mathbb{R}_m$. Il volume del cubo è naturalmente uguale a 1.

Poniamo $v_0(r) := 1$ e perciò anche $v_0 = 1$.

Osservazione 1.3. $v_1(r) = 2r$ *e quindi* $v_1 = 1$.

Nota 1.4. Denotiamo con Γ la *funzione gamma* che, come è noto dall'analisi, è in primo luogo un'interpolazione del fattoriale, che però appare in molti altri campi della matematica e deve essere considerata come la più importante funzione non elementare. Essa è definita e olomorfa su tutto il piano complesso tranne nei punti z della forma $z = -k$ con $k \in \mathbb{N}$ e soddisfa l'equazione funzionale

$$\Gamma(z+1) = z\Gamma(z)$$

per $z \in \mathbb{C} \setminus (-\mathbb{N})$. Vale la condizione iniziale $\Gamma(1) = 1$, da cui per induzione si ha come conseguenza immediata che

$$\Gamma(n+1) = n!$$

per ogni $n \in \mathbb{N}$. Si dimostra inoltre che $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Purtroppo, per pure ragioni storiche, la funzione è definita in modo tale che al fattoriale $n!$ non corrisponda l'argomento n in Γ.

Una trattazione molto dettagliata della funzione Γ si trova nel testo di analisi complessa di Remmert.

Teorema 1.5. *Il volume della palla unitaria in* $\mathbb{R}_m$ *è dato da*

$$v_m(1) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)}$$

Dimostrazione. Corsi di analisi. La formula è facile da ricordare nella forma

$$v_m(1) = \frac{\pi^{\frac{m}{2}}}{\frac{m}{2}!}$$

che è corretta per m pari e può essere considerata come abbreviazione simbolica nel caso che m sia dispari.

Proposizione 1.6. *Per* $m \geq 2$ *valgono le formule di ricorsione*

$$v_m(1) = \frac{2\pi}{m} v_{m-2}(1)$$
$$v_m(r) = \frac{2\pi r^2}{m} v_{m-2}(r)$$
$$v_m = \frac{\pi}{2m} v_{m-2}$$

Dimostrazione. (1) Abbiamo

$$v_m(1) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)} = \frac{\pi\pi^{\frac{m-2}{2}}}{\frac{m}{2}\Gamma(\frac{m}{2})} = \frac{\pi\pi^{\frac{m-2}{2}}}{\frac{m}{2}\Gamma(\frac{m-2}{2}+1)}$$
$$= \frac{2\pi}{m}\frac{\pi^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2}+1)} = \frac{2\pi}{m} v_{m-2}(1)$$

Ciò mostra la prima formula da cui si ottengono facilmente le altre due perché è chiaro che aumentando la dimensione di due il volume deve essere moltiplicato con r^2.

Corollario 1.7. $\lim_{m\to\infty} v_m(r) = 0$ *per ogni* r.

Dimostrazione. Per $m \geq 4\pi r^2$ si ha

$$v_m(r) = \frac{2\pi r^2}{m} v_{m-2}(r) \leq \frac{1}{2} v_{m-2}(r)$$

Possiamo scrivere una funzione in Python con cui calcoliamo i valori di v_m per $m \leq 20$. L'ultima cifra è arrotondata, le altre cifre sono corrette. Vediamo che per $m = 10$ la palla occupa solo il 2.49 per mille del volume del cubo a cui è iscritta!

```
def volume (m): # M.volume
  if m<=1: return 1
  else: return volume(m-2)*math.pi/(m+m)

for m in xrange(1,21): print '%2d   %.14f' %(m,M.volume(m))
```

m	v_m
1	1.00000000000000
2	0.78539816339745
3	0.52359877559830
4	0.30842513753404
5	0.16449340668482
6	0.08074551218828
7	0.03691223414321
8	0.01585434424382
9	0.00644240020066
10	0.00249039457019
11	0.00091997259736
12	0.00032599188693
13	0.00011116073667
14	0.00003657620418
15	0.00001164072512
16	0.00000359086045
17	0.00000107560049
18	0.00000031336169
19	0.00000008892365
20	0.00000002461137

Quale vicinanza?

Dalla tabella si vede che $v_{20} = 0.0000000246\ldots < 10^{-7}$. Se abbiamo quindi raccolto le concentrazioni nel sangue di 20 molecole rappresentate da numeri in $[0,1]$ di un milione di pazienti (un numero difficilmente raggiungibile nella realtà) e se volessimo considerare i dati x e y di due pazienti simili se $|x-y| < 0.5$ nella metrica euclidea di $\mathbb{R}_{20}$, la probabilità che per un punto x ce ne sia uno distinto e vicino (in questo senso) è solo circa 0.1 e quindi spesso questo concetto di vicinanza risulta poco utilizzabile.

La lunghezza della diagonale

Mentre il raggio della palla iscritta al cubo unitario in $\mathbb{R}_m$ è sempre uguale a $\frac{1}{2}$, il diametro del cubo, cioè la lunghezza della diagonale tra l'origine e il punto $(1,\ldots,1)$, è uguale a $\sqrt{m}$. Ciò implica che la palla, pur toccando il bordo del cubo (nei centri dei sottocubi di dimensione $m-1$), dista invece dai vertici di $\dfrac{\sqrt{m}-1}{2}$; siccome questa distanza diventa sempre più grande, ciò crea l'impressione che il cubo m-dimensionale al crescere di m assomiglia sempre di più a un riccio con corpo sferico sempre più piccolo e aculei sempre più lunghi.

Il problema del guscio

X sia un sottoinsieme misurabile di misura $v(X) < \infty$ in $\mathbb{R}_m$ e $0 \leq \alpha < 1$. Allora $v(\alpha X) = \alpha^m v(X)$. X sia *stellato* rispetto all'origine e quindi $\alpha X \subset X$. Per il volume del guscio $X \setminus \alpha X$ si ha allora

$$v(X \setminus \alpha X = (1 - \alpha^m) v(X)$$

e quindi

$$\frac{v(X \setminus \alpha X)}{v(X)} = 1 - \alpha^m$$

Siccome $\alpha < 1$ questo rapporto tende a 1; ciò significa che il guscio occupa, con il crescere di m, un volume relativo sempre maggiore.

Questo fenomeno è importante in statistica perché implica che in alta dimensione la maggior parte di una popolazione casuale si troverà in posizioni marginali dello spazio dei dati venendo così meno quanto si osserva nella statistica univariata in cui i valori di una popolazione *normale* si concentrano nella vicinanza del valore medio.

Il paradosso delle pareti

Già in dimensione 5 si verifica un fenomeno molto sorprendente impossibile in dimensioni ≤ 3 e probabilmente anche in dimensione 4. Troviamo infatti adesso una sfera in $\mathbb{R}_5$ che interseca tutti i lati 4-dimensionali del cubo unitario, ma non contiene il centro del cubo! Procediamo in questo modo:

Il centro del cubo è

$$c := (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$$

Il punto

$$p := (\alpha, \alpha, \alpha, \alpha, \alpha)$$

con $\alpha = 0.08$ appartiene anch'esso al cubo. Inoltre

$$|p - c|^2 = 5 \cdot (\alpha - \frac{1}{2})^2 = 5 \cdot 0.42^2 = 0.882$$

Se scegliamo il raggio $\rho = 0.93$, allora $\rho^2 = 0.8649$, per cui il centro c non appartiene alla palla di raggio ρ attorno a p. Facciamo adesso vedere che la sfera di raggio ρ interseca ogni lato 4-dimensionale del cubo.

Un tale lato è dato dall'intersezione del cubo con un iperpiano dato da un'equazione della forma $x_j = 0$ oppure $x_j = 1$. Per simmetria possiamo assumere che $j = 1$.

È suffciente dimostrare che esistono $\beta, \gamma \in [0, 1]$ tali che i punti

$$q_1 := (0, \beta, \beta, \alpha, \alpha)$$

e

$$q_2 := (1, \gamma, \alpha, \alpha, \alpha)$$

hanno distanza ρ da p. Abbiamo

$$|q_1 - p|^2 = \alpha^2 + 2(\beta - \alpha)^2$$

e

$$|q_2 - p|^2 = (1 - \alpha)^2 + (\gamma - \alpha)^2$$

(1) Per q dobbiamo soddisfare l'equazione

$$\rho^2 = \alpha + 2(\beta - \alpha)^2$$

cioè

$$\frac{\rho^2 - \alpha^2}{2} = (\beta - \alpha)^2$$

Ma

$$\frac{\rho^2 - \alpha^2}{2} = 0.42925$$

quindi bisogna avere

$$|\beta - \alpha| = \sqrt{0.42925}$$

per cui possiamo porre

$$\beta = \sqrt{0.42925} + \alpha$$
$$= 0.65517\ldots + 0.08 = 0.73517\ldots$$

Abbiamo quindi $\beta \in [0, 1]$.

(2) Per q_2 dobbiamo avere

$$\rho^2 = (1 - \alpha)^2 + (\gamma + \alpha)^2$$

cioè

$$\rho^2 - (1 - \alpha)^2 = (\gamma - \alpha)^2$$

Ma

$$\rho^2 - (1 - \alpha)^2 = 0.8649 - 0.92^2$$
$$= 0.8649 - 0.8464 = 0.0185$$

e quindi bisogna avere

$$|\gamma - \alpha| = \sqrt{0.0185}$$

per cui possiamo porre

$$\gamma = \sqrt{0.0185} + \alpha = 0.21601\ldots$$

Anche $\gamma \in [0, 1]$.

Modificato da Böhm/, pag. 6, in cui si dà un esempio per $m = 16$. Abbiamo chiamato questo esempio il paradosso delle pareti, perché per convincersi della stranezza dell'enunciato è sufficiente immaginare che una sfera possa intersecare tutte le pareti di una stanza cubica senza contenere il punto centrale della stanza.

Questi fenomeni creano molti problemi nell'interpretazione statistica e nello sviluppo degli algoritmi in alte dimensioni.

C. Böhm/S. Berchtold/D. Keim: Searching in high-dimensional spaces - index structures for improving the performance of multimedia databases. Internet ca. 2001, 74p.

Il paradosso della sfera centrale

Consideriamo un cubo stavolta con centro nell'origine di $\mathbb{R}_m$ e di lato 4. Nei 2^m punti della forma $(\pm 1, \ldots, \pm 1)$ in cui i segni $+$ o $-$ vengono scelti in tutti i modi possibili, poniamo una sfera di raggio 1 con centro in quel punto. Consideriamo poi la sfera con centro nell'origine tangente a tutte quelle altre sfere. Il suo raggio sia ρ. La situazione è illustrata per $m = 2$ dalla figura in alto nella colonna accanto.

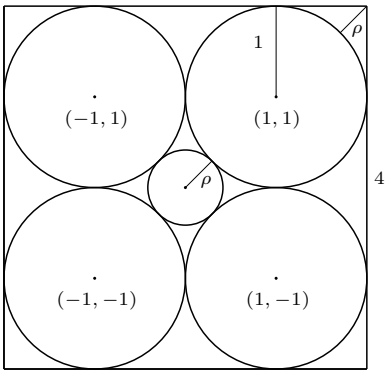
Il mezzo diametro del cubo è uguale a $2\sqrt{m}$, ma anche a $2\rho + 2$, abbiamo quindi

$$1 + \rho = \sqrt{m}$$

per cui

$$\rho = \sqrt{m} - 1$$

Ciò significa che per $m = 9$ la sfera interna tocca il bordo del cubo e per $m \geq 10$ esce addirittura da esse, benché le altre sfere rimangano naturalmente tutte contenute nel cubo. Da Gentle, pag. 297.



D. Donoho: High-dimensional data analysis - the curses and blessings of dimensionality. Internet 2000, 32p.

M. Gasparini: La statistica nelle prove cliniche. Boll. UMI Mat. Soc. Cult. 6/A (2003), 119-140.

J. Gentle: Elements of computational statistics. Springer 2002.

T. Hastie/R. Tibshirani/J. Friedman: The elements of statistical learning. Springer 2001.

R. Remmert: Classical topics in complex function theory. Springer 1998.

B. Schölkopf/A. Smola: Learning with kernels. MIT 2002. Tra le tecniche più popolari per superare i problemi dell'alta dimensione, gli algoritmi ai vettori di supporto si basano sulla teoria degli spazi di Hilbert con nucleo.

Proiezioni ottimali

La teoria delle proiezioni ottimali (che nella letteratura inglese appare sotto il nome di *projection pursuit*) è stata iniziata da Friedman e Tukey. Si cercano proiezioni ottimali rispetto a una funzione (*indice*) di rilevanza che può essere scelta in vari modi. Questo metodo interessante, piuttosto impegnativo nel calcolo, che, almeno nelle intenzioni, permette di superare le difficoltà delle alte dimensioni e che contiene come casi speciali e in un certo senso migliora molti metodi classici della statistica multivariata (come l'analisi delle componenti principali e l'analisi delle discriminanti) è esposto in un famoso articolo di Peter Huber e nella tesi di Guy Nason. Il pacchetto *XGobi* di R contiene funzioni per questa tecnica.

D. Cook/A. Buja/J. Cabrera/C. Hurley: Grand tour and projection pursuit. J. Comp. Graph. Stat. 4 (1995), 155-172.

P. Huber: Projection pursuit. Ann. Statistics 13 (1985), 435-475.

G. Nason: Design and choice of projection indices. PhD thesis Bath Univ. 1992.

www.stats.bris.ac.uk/ guy/. Sito di Guy Nason a Bristol. Seguendo la voce *Research* si trovano tra l'altro la sua tesi e software riguardanti il metodo delle proiezioni ottimali.

www.ggobi.org. GGobi è il successore di XGobi.