

Data science con MATLAB

Marco Riani¹, Dipartimento di Scienze Economiche e Aziendale
and Interdepartmental Research Centre for Robust Statistics,
Università di Parma, 43100 Parma, Italy

25 novembre 2021

¹e-mail: mrhani@unipr.it

Indice

1	Algebra lineare	9
1.1	Operazioni elementari con le matrici	9
1.1.1	La generazione di numeri casuali	12
1.1.2	Le matrici diagonali	16
1.1.3	Le matrici idempotenti	21
1.1.4	Lo spazio vettoriale	22
1.1.5	La norma di un vettore	22
1.1.6	Il prodotto scalare	27
1.1.7	Le forme quadratiche	32
1.1.8	Estrazione degli elementi tramite forme quadratiche . .	41
1.1.9	Matrici ortogonali	43
1.1.10	Moltiplicazione di matrici trasposte	44
1.1.11	Moltiplicazione di matrici inverse	46
1.1.12	La trasposta dell'inversa	47
1.1.13	La traccia	48
1.1.14	Dipendenza, indipendenza lineare e base di uno spazio vettoriale	50
1.1.15	Il rango	59

1.1.16	Sistemi di equazioni lineari	62
1.2	Autovalori e autovettori	64
1.2.1	Polinomio caratteristico	65
1.3	Routines per il calcolo degli autovalori e degli autovettori . . .	69
1.4	Scomposizione spettrale	70
1.4.1	La scomposizione spettrale attraverso il calcolo simbolico	72
1.5	Introduzione ai poligoni	74
1.6	Proiezioni ortogonali	83
1.7	L'espansione implicita	88
1.8	Matrice di varianze e correlazione tramite espressioni matriciali	93
2	Le distanze e gli indici di similarità	101
2.1	Definizione di distanze	102
2.2	Alcuni tipi di distanza	103
2.3	Gli indici di distanza e gli indici di dissimilarità	114
2.4	Lo spazio euclideo ponderato	116
2.5	La distanza di Mahalanobis	119
2.5.1	Proprietà della distanza di Mahalanobis	126
2.6	La scala di misura delle distanze	128
2.7	Gli indici di similarità	130
2.7.1	Indici di similarità per fenomeni dicotomici	131
2.7.2	Indici di similarità in presenza di fenomeni misti	139
3	La riduzione delle dimensioni	145
3.1	Analisi in componenti principali (PC): introduzione	145
3.2	La prima PC come combinazione lineare delle variabili originarie	147

3.3	Le prime k PC come combinazioni lineari delle variabili originarie	151
3.3.1	Relazione tra autovalori traccia e determinante	152
3.4	La scomposizione in valori singolari (svd)	154
3.5	Le prime k PC come migliore rappresentazione di rango k delle variabili originarie	158
3.6	PC come proiezione ortogonale dei punti in un sottospazio di dimensione ridotta	162
3.6.1	Retta di regressione e retta associata alla prima componente principale	164
3.6.2	Ricostruzione della matrice originaria con una matrice di rango ridotto	170
3.6.3	Componenti principali come rotazione degli assi cartesiani	174
3.7	L'analisi in componenti principali in pratica	186
3.8	Il biplot	202
3.9	Componenti principali su \tilde{X} oppure su Z	214
4	L'analisi delle corrispondenze	223
4.1	Notazione	229
4.2	Giudizi sulla bontà dell'analisi e punteggi	245
4.3	Contributi all'inerzia del punto o all'inerzia della dimensione latente	252

Elenco delle tabelle

- 1.1 Esempi di generazione di matrici definite positive, semidefinite e definite negative. 37
- 1.2 Spesa pubblicitaria e fatturato per 25 aziende del settore tessile (dati in milioni di Euro) 70
- 2.1 Matrice dei dati di partenza X . Numero di certificazioni ISO 9000 in 5 paesi in un periodo 75
- 2.2 Matrice dei dati di partenza X . Numero di ordini e ammontare (in migliaia di Euro) effettuati 76
- 2.3 Caratteristiche qualitative dicotomiche (prime 5 variabili) e quantitative (ultime tre variabili) 77
- 3.1 Spesa pubblicitaria e fatturato per 25 aziende del settore tessile (dati in milioni di Euro) nel 1995 80
- 3.2 Matrice dei dati riferita a 16 lavatrici. Sono state rilevate 6 variabili 205
- 4.1 Tabella di contingenza tra il grado di scolarizzazione e la posizione verso la scienza per un campione di 1000 persone 210
- 4.2 Tabella di contingenza tra il grado di istruzione e la professione lavorativa in un campione di 1000 persone 211
- 4.3 Tabella di contingenza relativa all'appartenenza al partito politico e alla posizione sulla pena di morte 212
- 4.4 Dati sull'importazione di generi di abbigliamento nella UE. Tabella di contingenza tra i 28 paesi 213
- 4.5 Tipologia di marca di dentifricio prevalentemente utilizzata da 1576 consumatori appartenenti a 16 fasce di reddito 214
- 4.6 Matrice dei profili riga ($R_{3 \times 4}$) della Tabella 4.5. In questa matrice la somma di ogni riga è uguale a 1 215
- 4.7 Matrice trasposta dei profili colonna ($C_{4 \times 3}$) della Tabella 4.5. In questa matrice la somma di ogni colonna è uguale a 1 216
- 4.8 Distanze al quadrato di ogni profilo riga dal profilo medio $\chi^2 d_{ic}^2$ e masse di riga (r). Distanze al quadrato di ogni profilo colonna dal profilo medio $\chi^2 d_{jc}^2$ e masse di colonna (c) 217
- 4.9 I profili riga e colonna definiscono due nuvole di punti nello spazio Euclideo ponderato di dimensione 3 218

4.10 Scomposizione dell'inerzia totale nelle k dimensioni latenti. La somma di ogni riga

Notazione

X = matrice dei dati originale con n righe e p colonne. Quando la dimensione della matrice è di particolare importanza viene indicata con $X_{n \times p}$. X può essere scritta come insieme di p vettori colonna X_1, \dots, X_p . X_j , vettore di dimensione di dimensione $n \times 1$ che contiene i valori (modalità) della variabile j , con $j = 1, 2, \dots, p$. $X = (X_1, \dots, X_j, \dots, X_p)$ dove X_j è il vettore colonna di lunghezza n definito come segue

$$\begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

X può anche essere scritta come insieme di n vettori riga $x'_1 \dots x'_n$ dove $x'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$.

$$X = \begin{pmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{pmatrix}$$

$\tilde{X}_{n \times p}$ = matrice degli scostamenti dalla media (la somma (media) di ogni colonna in questa matrice è pari a 0).

$S_{p \times p}$ = matrice di covarianze. Il generico elemento i, i lungo la diagonale principale di questa matrice è $var(X_i) = s_i^2$, $i = 1, 2, \dots, p$. Il generico elemento i, j di questa matrice è $cov(X_i, X_j)$, $i = 1, 2, \dots, p$, $j = 1, 2, \dots, p$.

$Z_{n \times p}$ = matrice che contiene gli scostamenti standardizzati. La somma

(media) di ogni colonna di questa matrice è 0 e la varianza di ogni colonna è pari ad 1. $Z = (Z_1, \dots, Z_p)$. $var(Z_j) = 1$,

$R_{p \times p}$ =matrice di correlazione. Il generico elemento i, i lungo la diagonale principale di questa matrice è 1, $i = 1, 2, \dots, p$. Il generico elemento i, j di questa matrice è $r(X_i, X_j)$ ossia la correlazione tra la variabile X_i e la variabile X_j .

$$r(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}}$$

Capitolo 1

Algebra lineare

1.1 Operazioni elementari con le matrici

Una matrice A è un insieme di $n \times p$ elementi disposti in n righe e p colonne. L'elemento situato all'incrocio tra la riga i e la colonna j viene indicato con a_{ij} . Ciascuna colonna (contenente n elementi) può essere vista come un punto nello spazio R^n . Viceversa, ciascuna delle n righe (contenente p elementi) può essere vista come un punto nello spazio R^p . Una matrice $n \times p$ è anche un insieme di p vettori colonna (oppure di n vettori riga). Si chiama scalare un vettore di un solo elemento.

Una matrice A di dimensione $n \times p$ si dice quadrata se $n = p$. Data una matrice A di dimensione $n \times p$ la sua trasposta (di dimensione $p \times n$) si ottiene da A scambiando tra loro le righe con le colonne. Se $A' = A$ allora la matrice si dice simmetrica.

Se c è uno scalare uguale e A e B sono due matrici le operazioni più comuni sono:

1. $c * A$ (moltiplicazione di uno scalare per una matrice);
2. A' (trasposta di A);
3. $A + B$ (somma elemento per elemento);
4. $A * B$ (moltiplicazione matriciale supponendo che le matrici siano conformabili. Date due matrici, se il numero di righe della seconda è uguale al numero di colonne della prima si dice che le matrici sono conformabili rispetto alla moltiplicazione infatti per poter effettuare la moltiplicazione di matrici esse devono essere tra loro conformabili. Il risultato sarà una matrice avente un numero di righe pari a quello del primo fattore (matrice premoltiplicanda) e un numero di colonne pari a quelle del secondo fattore (matrice postmoltiplicanda).

Se A è una matrice $m \times n$ e B è una matrice $n \times p$, $C = A \times B$ sarà una matrice $m \times p$)

Vediamo nel dettaglio come funziona la moltiplicazione fra matrici.

Supponiamo che A sia una matrice $m \times n$ e B una matrice $n \times p$:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}$$

Il risultato del prodotto AB è una matrice C di dimensioni $m \times p$, così fatta:

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{pmatrix}$$

dove ogni elemento c_{ij} della matrice C è costituito dalla sommatoria dei prodotti dei vettori riga della matrice A e dei vettori colonna della matrice B come segue:

$$c_{ij} = \sum_{k=1}^n a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{in}b_{nj}.$$

In questa ultima sommatoria risiede appunto il vincolo di conformabilità: il numero elementi moltiplicandi di A e di B deve essere per entrambe le matrici di lunghezza n , ossia che le righe di A siano di egual numero delle colonne di B .

Supponiamo che le matrici A e B siano così fatte:

$$A = \begin{pmatrix} 3 & -3 & 9 \\ 10 & 0 & 4 \\ 0 & 2 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 5 \\ 7 & 3 \\ 0 & 6 \end{pmatrix}$$

Dato che A è una matrice 3×3 e B è una matrice 3×2 , il vincolo di conformabilità è soddisfatto e il prodotto $C = A \times B$ risulta essere una

matrice C di dimensioni 3×2 :

$$\begin{pmatrix} 3 & -3 & 9 \\ 10 & 0 & 4 \\ 0 & 2 & -1 \end{pmatrix} \begin{pmatrix} -1 & 5 \\ 7 & 3 \\ 0 & 6 \end{pmatrix} = \begin{pmatrix} -24 & * \\ * & * \\ * & * \end{pmatrix}$$

L'elemento $c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} = 3 \cdot (-1) - 3 \cdot 7 + 9 \cdot 0 = -24$.

E' immediatamente evidente che non è possibile calcolare il prodotto $D = B \times A$ in quanto le due matrici non sono conformabili: non è possibile moltiplicare una matrice B con dimensioni 3×2 con una matrice A di dimensioni 3×3 , la proprietà commutativa del prodotto di scalari non si estende quindi al prodotto fra matrici e quindi $A \times B \neq B \times A$.

1.1.1 La generazione di numeri casuali

Tutte le funzioni di MATLAB che terminino con `rnd` generano matrici di numeri casuali da specifiche distribuzioni.

Ad esempio, la funzione `normrnd(2,3)` genera un numero casuale da una distribuzione normale con media (μ) pari a 2 (primo argomento di input) e scostamento quadratico medio (σ) pari a 3 (secondo argomento di input). Se la funzione `normrnd` viene chiamata con 4 argomenti di input come segue `normrnd(2,3,5,6)`, è possibile specificare quante righe e quanto colonne deve avere la matrice di numeri casuali. Ad esempio `normrnd(10,1,5,6)` produce una matrice di dimensione 5×6 di numeri casuali da una distribuzione normale con media 10 e $\sigma = 1$.

Similmente, la funzione `chi2rnd(5,2,3)` genera una matrice di dimensione 2×3 di numeri casuali da una distribuzione χ^2 con 5 gradi di libertà.

Le funzioni `rand` e `randn` generano matrici rispettivamente da una distribuzione uniforme con valori nell'intervallo $[0; 1]$ e da una distribuzione normale standardizzata ossia con $\mu = 0$ e $\sigma = 1$ ($N(0,1)$). Ad esempio `randn(3,4)` genera una matrice 3×4 di numeri casuali dalla distribuzione normale standardizzata.

La funzione `randi` consente di ottenere numeri casuali interi uniformemente distribuiti nell'intervallo specificato. Ad esempio `randi([2 10],3,4)` restituisce una matrice di dimensione 3×4 di numeri casuali interi nell'intervallo $[2 \ 10]$,

Osservazione: la demo `randtool` consente di generare numeri casuali da un'ampia gamma di distribuzioni in maniera interattiva.

Osservazione: finora abbiamo parlato di numeri casuali anche se in realtà sarebbe stato più corretto parlare di numeri pseudocasuali, poiché i numeri sono generati da un algoritmo deterministico che produce una sequenza. Questo algoritmo può essere inizializzato con un seme (seed). Per visualizzare l'attuale valore del seed si può digitare dal prompt `rng`. Normalmente all'avvio MATLAB setta il seed a 0.

```
>> rng
```

```
ans =
```

```
struct with fields:
```

```
Type: 'twister'
```

```
Seed: 0
```

```
State: [625 x 1 uint32]
```

Il seed può essere modificato con l'istruzione `rng`(numero naturale a piacere minore di 2^{32}).

Ad esempio, le istruzioni

```
>> rng(100)
```

```
>> randn(2,3)
```

producono sempre i numeri casuali di seguito

```
ans =
```

```
0.1609    -0.2390    -0.9014
-0.6151     0.6150     0.3481
```

Esercizio: generare una matrice di dimensione 1000×2 di numeri casuali dalla distribuzione normale multivariata con vettore delle medie $\mu = (2 \ 5)$ e matrice di covarianze

$$\Sigma = \begin{pmatrix} 3 & 1.6 \\ 1.6 & 1.5 \end{pmatrix}$$

Per la replicabilità dei risultati utilizzare il seed 234. Tramite la funzione `scatterboxplot` fare il grafico dei dati che sono stati generati.

Soluzione

```
rng(234)
```

```
n=1000;
```

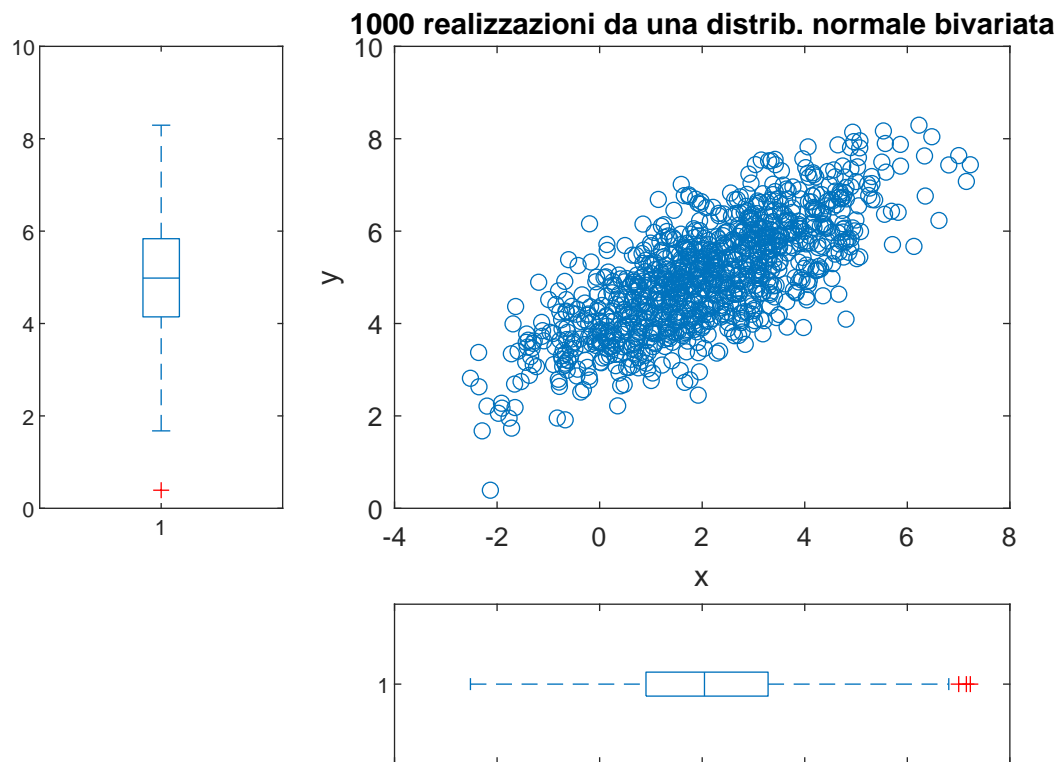



Figura 1.1: 1000 realizzazioni da una distribuzione normale bivariata con vettore delle medie $\mu = (2 \ 5)'$ e matrice di covarianze specificata con correlazione positiva

```
mu=[2 5];
sigma12=1.6;
Sigma=[ 3,sigma12; sigma12 , 1.5];
X=mvnrnd(mu,Sigma,n);
scatterboxplot(X(:,1),X(:,2))
title('1000 realizzazioni da una distrib. normale bivariata')
```

Il grafico che appare è riportato nella Figura 1.1.

1.1.2 Le matrici diagonali

La matrice diagonale è una matrice in cui solamente i valori della diagonale principale possono essere diversi da 0. Solitamente per la matrice diagonale D di ordine n si utilizza la notazione $\text{diag}(d_1, d_2, \dots, d_n)$ per indicare i valori d_1, d_2, \dots, d_n posti in sequenza sulla diagonale principale (a partire dall'angolo superiore sinistro).

Esercizi sulle matrici diagonali.

Creare la matrice $D = \text{diag}(2, 5, 7)$.

Soluzione:

```
>> D=diag([2 5 7])
```

Output

D =

$$\begin{array}{ccc} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{array}$$

Creare la matrice E che segue (valori non nulli sopra la diagonale principale)

$$E = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 7 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Soluzione

```
>> E=diag([2 5 7],1)
```

Se la funzione `diag` viene chiamata con due argomenti di input allora il secondo argomento di input (numero intero k) specifica su quale diagonale inserire i numeri. $k = 0$, inserisce i numeri sulla diagonale principale, $k = 1$ (come nell'esempio precedente) inserisce i numeri sopra la diagonale principale, ecc. ... $k = -2$ inserisce i numeri due linee sotto la diagonale principale.

Osservazione: occorre prestare attenzione al fatto che la dimensione della matrice di output dipende dal valore di k .

Esercizio.

Data la seguente matrice A definita come segue

$$A = \begin{pmatrix} 12 & 4 & 11 & 10 & 9 \\ 1 & 3 & 15 & 11 & 3 \\ 10 & 12 & 1 & 5 & 6 \\ 12 & 3 & 8 & 14 & 11 \\ 8 & 2 & 13 & 11 & 7 \end{pmatrix}$$

estrarre gli elementi sulla diagonale principale in un vettore denominato a e gli elementi di 3 linee sotto la diagonale principale in un vettore denominato b .

Osservazione: la matrice di cui sopra è stata generata dalle seguenti istruzioni

```
rng(10); A=randi(15,5)
```

Soluzione

Per estrarre gli elementi sulla diagonale principale si chiama la funzione `diag`

in uno dei due modi che seguono:

$$a = \text{diag}(A) \quad a = \text{diag}(A, 0)$$

$a =$

12

3

1

14

7

Per estrarre gli elementi di due linee sotto la diagonale principale si chiama la funzione `diag` con il secondo argomento pari a -2 come segue

$$b = \text{diag}(A, -2)$$

$b =$

10

3

13

Matrice identità = una matrice quadrata in cui tutti gli elementi della diagonale principale sono costituiti dal numero 1, mentre i restanti elementi sono 0. Viene indicata con I oppure con I_n , dove n è il numero di righe (colonne) della matrice.

La proprietà fondamentale di I_n è che: $AI_n = A$ e $I_nB = B$ per ogni matrice A e B per cui sono definite queste moltiplicazioni di matrici.

Esercizi sulle matrici identità.

Creare la matrice identità I_3

Soluzione

`I=eye(3)`

`I =`

1	0	0
0	1	0
0	0	1

Se la funzione `eye` viene chiamata con 2 argomenti di input (r, c) , oppure con un solo argomento di input di lunghezza 2 (che specifica il numero di righe e di colonne `[4,6]`), viene creata una matrice $r \times c$ con elementi pari ad 1 sulla diagonale principale e pari a 0 fuori dalla diagonale.

Esercizio

Creare la matrice che segue

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Soluzione

`A=eye(4,6);` oppure `A=eye([4,6]);`

Esercizio: creare la matrice identità di ordine 5 tramite doppio ciclo for.

Soluzione

```
p=5;
id=zeros(p,p);
for i=1:p
    for j=1:p
        if i==j
            id(i,i)=1;
        end
    end
end
I=eye(p);

assert(isequal(I,id),"errore di programmazione nella " + ...
    "creazione della matrice identità tramite doppio ciclo for")
```

In MATLAB le funzioni `zeros` e `ones` consentono di ottenere rispettivamente matrici con elementi tutti uguali a 0 oppure a 1 e seguono la stessa sintassi di `eye`. Ad esempio

```
>> ones(2,3)
```

```
ans =
```

```
1     1     1
1     1     1
```

```
>> zeros(3,4)
```

ans =

```

0     0     0     0
0     0     0     0
0     0     0     0

```

In questo testo per indicare una matrice $n \times p$ composta da elementi tutti uguali ad 1 si usa il simbolo $1_{n \times p}$. Secondo questa simbologia, quindi $1_{5 \times 1}$ indica un vettore colonna di lunghezza 5 con tutti gli elementi uguali ad 1 (l'istruzione MATLAB è `ones(5,1)`). Per indicare una matrice $n \times p$ composta da elementi tutti uguali ad 1 si usa il simbolo $1_{n \times p}$.

Esercizio:

Generare una matrice di dimensione 3×4 con elementi tutti uguali a 5.

Soluzione: `5*ones(3,4)`

1.1.3 Le matrici idempotenti

Una matrice quadrata H si dice idempotente se $H \times H = H$.

Esercizi sulle matrici idempotenti:

Inserire un valore numerico intero positivo in una variabile denominata n . Creare, tramite la funzione `gallery`¹, la matrice denominata `'involv'`. Aggiungere a questa matrice la matrice identità e dividere il risultato per due. Chiamare la matrice risultante H . Verificare che H è idempotente

Soluzione

¹La funzione `gallery` contiene oltre 50 matrici speciali formulate per tecniche di verifica numerica.

```
% Assegno ad n il valore 5.
n=5;
A= gallery('invol',n);
H=0.5*(eye(n)+A);
disp('Esempio di matrice idempotente')
disp(H)
H2=H*H;
disp('Verifica idempotenza di H')
maxdiff=max(abs(H-H2),[],'all')<1e-10;
assert(maxdiff,"Errore di programmazione nella verifica ..." + ...
        "dell'idempotenza di H")
```

1.1.4 Lo spazio vettoriale

Uno spazio vettoriale è una struttura algebrica composta da un insieme di scalari (detto campo²), un insieme di vettori e da due operazioni binarie (somma e moltiplicazione per uno scalare) caratterizzate da determinate proprietà. Ad esempio, l'insieme delle matrici A di dimensioni $m \times n$ con le operazioni di somma tra matrici e prodotto di uno scalare per una matrice, è uno spazio vettoriale.

1.1.5 La norma di un vettore

La norma $\|x\|_2 = \|x\|$ (Euclidea) di un vettore $x = (x_1, x_2, \dots, x_n)'$ di dimensione n (anche detta modulo o lunghezza di x) è definita dalla seguente

²I due campi più noti sono l'insiemi dei numeri reali \mathbb{R} oppure l'insieme dei numeri complessi \mathbb{C} .

espressione.

$$\|x\|_2 = \|x\| = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x'x} \quad (1.1)$$

La norma Euclidea di un vettore, quindi, si calcola estraendo la radice quadrata della somma dei quadrati delle componenti del vettore.

È immediato osservare che nella matrice degli scostamenti dalla media \tilde{X} le norme dei vettori colonna $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$ non sono altro che gli scostamenti quadratici medi delle variabili originarie moltiplicati per $\sqrt{n-1}$.

Nella matrice degli scostamenti standardizzati Z le norme dei vettori colonna Z_1, Z_2, \dots, Z_p sono tutte uguali a $\sqrt{n-1}$. Dal punto di vista algebrico quindi, l'operazione di standardizzazione equivale a lavorare con vettori colonna (variabili) che hanno la stessa origine e presentano la stessa norma.

Esercizio.

Generare una matrice di numeri casuali di dimensione $n \times p$ dalla distribuzione χ^2 con 5 gradi di libertà.

1. Verificare tramite un ciclo **for** che la norma al quadrato di ogni colonna della matrice degli scostamenti dalla media divisa per $\sqrt{n-1}$ non è altro che la varianza campionaria delle variabili originarie. Calcolare la norma manualmente e tramite la funzione di MATLAB denominata **norm**
2. Tenendo presente che i numeri nella matrice sono stati generati da una distribuzione χ^2 con 5 gradi di libertà, quali valori ci attendiamo per

le medie e le varianze campionarie di ogni colonna?

3. Verificare tramite un ciclo **for** che la norma di ogni colonna della matrice degli scostamenti standardizzati divisa per $\sqrt{n-1}$ è pari ad 1. Calcolare la norma manualmente e tramite la funzione di MATLAB denominata **norm**. Inserire gli **assert** per controllare l'uguaglianza delle diverse implementazioni.

Per verificare la convergenza dei valori campionari ai rispettivi valori teorici, è opportuno scegliere n molto grande. In questo esempio poniamo $n = 100000$ e $p = 3$. Per la replicabilità dei risultati, fissiamo il seed dei numeri casuali a 25.

Soluzione

```
rng(25)
n=100000;
p=3;
X=chi2rnd(5,n,p);
medie=mean(X);
varianze=var(X);
sigma=std(X);

% Osservazione: dato che la v.c. Chi2 con g gradi di libertà ha una
% expectation pari a g e una varianza pari a 2g ci attendiamo che le medie
% campionarie siano vicine a 5 e che le varianze campionarie siano vicine a
```

```
% 10.

disp("Medie campionarie di variabili generate da Chi2(5)")
disp(medie)

disp("Varianze campionarie di variabili generate da Chi2(5)")
disp(varianze)


Xtilde=X-medie;
Z=zscore(X);

%% Calcolo delle norme richieste tramite ciclo for
for j=1:p
    Xtildej=Xtilde(:,j);
    normXtildej=sqrt(sum(Xtildej.^2));
    % Di seguito chiamiamo direttamente la funzione norm
    normXtildejCHK=norm(Xtildej);
    diff=abs(normXtildej-normXtildejCHK);
    assert(diff<1e-12,"Errore di programmazione " + ...
        "nell'implementazione della norma euclidea")
    diff1=abs(normXtildej-sigma(j)*sqrt(n-1));
    assert(diff1<1e-12,"Errore di programmazione " + ...
        "nell'implementazione della norma euclidea")
    Zj=Z(:,j);
    normZj=sqrt(sum(Zj.^2));
    % Di seguito chiamiamo direttamente la funzione norm
    normZjCHK=norm(Zj);
    diff2=abs(normZj-normZjCHK);
```

```

assert(diff2<1e-12,"Errore di programmazione " + ...
      "nell'implementazione della norma euclidea")
diff3=abs(normZj-sqrt(n-1));
assert(diff3<1e-12,"Errore di programmazione " + ...
      "nell'implementazione della norma euclidea")
end

```

Nell'esercizio di cui sopra abbiamo chiamato la funzione `norm` con un solo argomento di input. Se viene specificato il secondo argomento di input allora è possibile stabilire la potenza a cui devono essere elevate le componenti del vettore. In generale, la norma di un vettore x di ordine k è definita da

$$||x||_k = \left(\sum_{i=1}^n |x_i|^k \right)^{\frac{1}{k}}. \quad (1.2)$$

Nel caso di $k = 1$ si fa la somma dei valori assoluti. La norma euclidea che abbiamo visto in precedenza si ottiene ponendo $k = 2$. Se $k \rightarrow \infty$ allora si ottiene il massimo tra i valori assoluti delle singole componenti. Similmente, se $k \rightarrow -\infty$ si ottiene il minimo tra i valori assoluti delle singole componenti.

Esercizio. Dato il vettore $[-3 \ 2 \ -10 \ 5 \ 20]$, calcolare le norme $-\infty$, 1 , 2 , ∞ , Inserire i 4 risultati dentro una table con intestazioni di righe e colonna appropriati.

Soluzione

```

x=[-3 2 -10 5 20];
% pp vettore che contiene le norme da calcolare
pp=[-Inf; 1; 2; Inf];

```

```

% norme= vettore che conterrà i risultati del calcolo
norme=zeros(length(pp),1);

for i=1:length(pp)
    norme(i)=norm(x,pp(i));
end

% lab=etichette da inserire come nomi di riga nella table
lab="Norma p="+ string(pp);
normeT=array2table(norme,"RowNames",lab,"VariableNames","Norme");
disp(normeT)

```

Lo script di cui sopra produce

	Norme

Norma p=-Inf	2
Norma p=1	40
Norma p=2	23.195
Norma p=Inf	20

1.1.6 Il prodotto scalare

Dati due vettori x e y di dimensione n , il numero dato dalla somma dei prodotti delle componenti corrispondenti dei due vettori

$$\sum_{i=1}^n x_i y_i = x' y$$

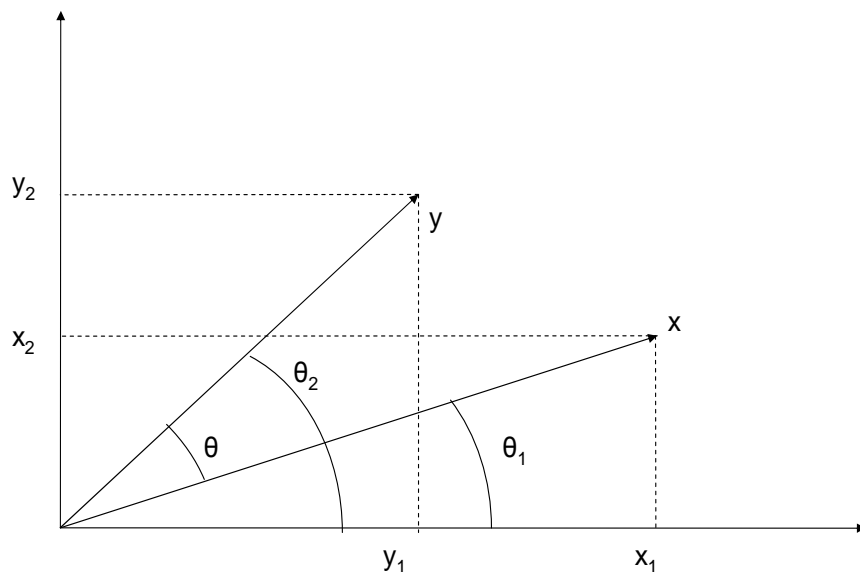


Figura 1.2: L'angolo θ fra $x = (x_1, x_2)'$ e $y = (y_1, y_2)'$ è il prodotto scalare tra i due vettori diviso per le rispettive norme.

viene chiamato prodotto interno o prodotto scalare.

Il coseno dell'angolo θ tra due vettori $x = (x_1, x_2)'$ e $y = (y_1, y_2)'$ come mostrato nella Figura 1.2 è dato dal prodotto scalare dei due vettori diviso dalle due norme.

$$\cos(\theta) = \frac{x_1 y_1 + x_2 y_2}{\|x\| \|y\|}. \quad (1.3)$$

Dimostrazione

Dalle definizioni di coseno e seno, utilizzando la Figura 1.2, si ha, $\cos(\theta_1) = x_1/\|x\|$ e $\cos(\theta_2) = y_1/\|y\|$, $\sin(\theta_1) = x_2/\|x\|$ e $\sin(\theta_2) = y_2/\|y\|$.

Ricordando che

$$\cos(\theta) = \cos(\theta_2 - \theta_1) = \cos(\theta_2) \cos(\theta_1) + \sin(\theta_2) \sin(\theta_1). \quad (1.4)$$

possiamo scrivere

$$\cos(\theta) = \cos(\theta_2 - \theta_1) = \frac{y_1}{\|y\|} \frac{x_1}{\|x\|} + \frac{y_2}{\|y\|} \frac{x_2}{\|x\|} = \frac{x_1 y_1 + x_2 y_2}{\|x\| \|y\|} = \frac{x' y}{\|x\| \|y\|}. \quad (1.5)$$

Poiché $\cos(90^\circ) = \cos(270^\circ) = 0$ e $\cos(\theta) = 0$ solo se $x' y = 0$, x e y sono perpendicolari quando $x' y = 0$.

Se i due vettori x e y hanno norma unitaria, il prodotto interno, quindi, non è altro che il coseno dell'angolo fra essi.

Due vettori colonna di lunghezza n si dicono ortogonali (perpendicolari) se il loro prodotto interno $x' y = 0$. Ad esempio, è facile verificare che i due vettori colonna v_1 e v_2 definiti come $v_1 = (a_1, 0)'$ e $v_2 = (0, a_2)'$ dove a_1 e a_2 sono due numeri reali qualsiasi sono perpendicolari in quanto $v_1' v_2 = 0$.

Esercizio

Dati due vettori $v_1 = (a, b)'$ e $v_2 = (b, -a)'$, rappresentare questi due vettori utilizzando la funzione `quiver`. Trovare l'angolo θ (in gradi) tra i due vettori e stampare il suo valore. Trovare i coefficienti angolari (m_1 e m_2) delle rette che passano attraverso questi due vettori. Verificare la condizione di perpendicolarità tra le due rette³. Per replicabilità dei risultati porre $a = 0.9796$; e $b = 0.2011$;

Soluzione

```
a=0.9796;
```

```
b=0.2011;
```

```
v1=[a;b];
```

³Due rette si dicono perpendicolari se e solo se i coefficienti angolari sono uno il reciproco dell'opposto dell'altro
<https://en.wikipedia.org/wiki/Perpendicular>.

```
v2=[b;-a];  
  
% Disegno il primo vettore  
quiver(0,0,a,b)  
  
% hold('on') per fare in modo che il grafico successivo si sovrapponga al  
% grafico precedente  
hold('on')  
quiver(0,0,b,-a)  
  
  
% Creo una nuova figura  
figure  
  
% Faccio una sola chiamata alla funzione quiver  
zer=zeros(2,1);  
quiver(zer,zer,v1,v2)  
  
% La funzione acosd ritorna la funzione inversa del coseno in gradi  
theta=acosd(v1'*v2);  
  
disp(["L'angolo tra i due vettori è " string(theta) "gradi"])  
  
% m1=coefficiente angolare della retta che passa  
% attraverso il primo vettore  
m1=v1(2)/v1(1);  
  
% m2=coefficiente angolare della retta che passa  
% attraverso il secondo vettore  
m2=v2(2)/v2(1);  
  
% Verifico che m1=-1/m2  
  
% https://it.wikipedia.org/wiki/Perpendicolarità  
disp(['m1=' num2str(m1)])
```



```
disp(['-1/m2=' num2str(-1/m2)])
```

Il codice di cui sopra produce la Figura 1.3.

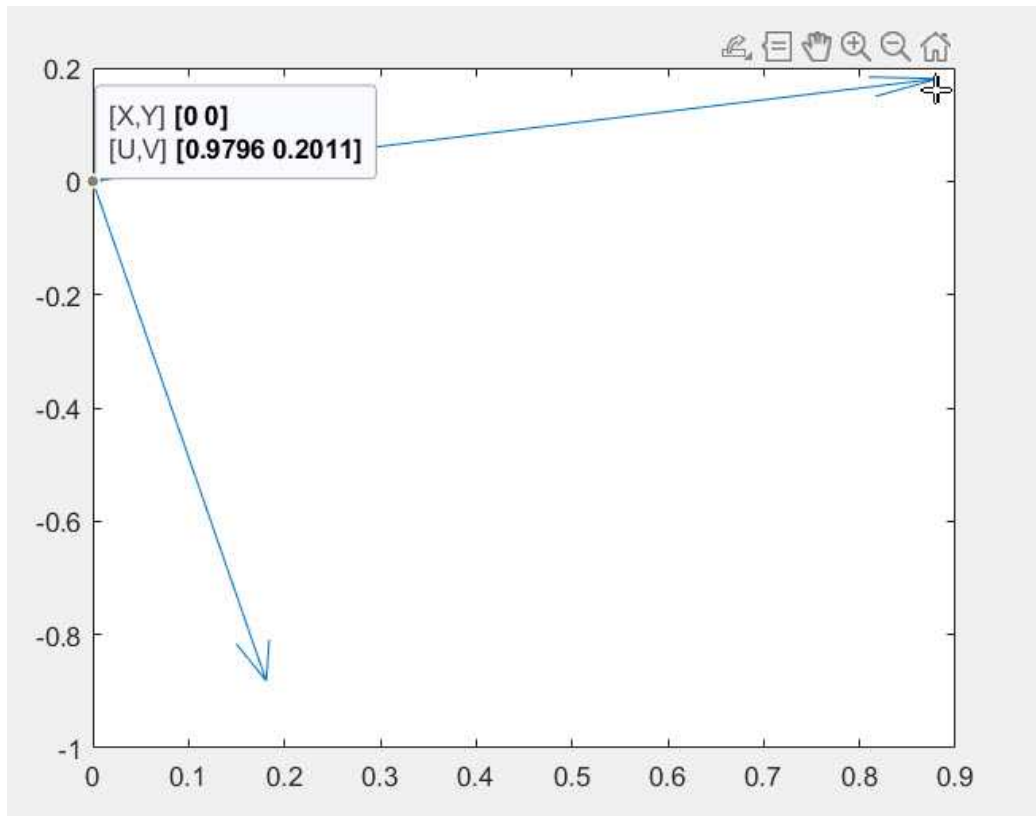


Figura 1.3: Esempio di vettori ortogonali. Facendo click su un punto qualsiasi di ciascuna freccia è possibile visualizzare un riquadro che contiene le coordinate di inizio e fine del vettore.

Si chiama vettore elementare e_i un vettore di n elementi tutti nulli tranne l' i -esimo che assume valore 1. Ad esempio se $n = 5$; $e'_2 = (0 \ 1 \ 0 \ 0 \ 0)$, $e'_5 = (0 \ 0 \ 0 \ 0 \ 1)$. Osserviamo che i vettori elementari e_1, e_2, \dots, e_n sono a due a due ortogonali $e'_i e_j = 0$ per $i \neq j$.

Dall'equazione (1.2) segue immediatamente che se i vettori x e y sono due generiche colonne della matrice degli scostamenti dalla media \tilde{X} , oppure

della matrice Z , il coseno dell'angolo tra questi due vettori non è altro che il coefficiente di correlazione lineare tra le variabili originarie.

In simboli MATLAB il coefficiente di correlazione lineare tra la colonna i e j della matrice dei dati r_{X_i, X_j} si può scrivere come

$$\text{Xtilde}(:,i)' * \text{Xtilde}(:,j) / (\text{norm}(\text{Xtilde}(:,i)) * \text{norm}(\text{Xtilde}(:,j)))$$

dove si suppone che la variabile Xtilde contenga \tilde{X} , oppure

$$Z(:,i)' * Z(:,j) / (\text{norm}(Z(:,i)) * \text{norm}(Z(:,j))) = Z(:,i)' * Z(:,j) / (n-1)$$

dove $n = \text{size}(Z, 1)$ è la numerosità campionaria e si suppone che la variabile Z contenga la matrice degli scostamenti standardizzati.

1.1.7 Le forme quadratiche

Nell'equazione (1.1) abbiamo che visto che la somma dei quadrati di un vettore colonna x può essere scritto come il vettore trasposto per se stesso.

$$\sum_{i=1}^n x_i^2 = x'x$$

La somma dei quadrati ponderati $\sum_{i=1}^n \lambda_i x_i^2$ può essere scritta come

$$\begin{aligned} \sum_{i=1}^n \lambda_i x_i^2 &= x' \Lambda x = (x_1 \ x_2 \ \dots \ x_n) \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \\ &= x' \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) x \end{aligned}$$

Si definisce forma quadratica una matrice quadrata postmoltiplicata per un vettore colonna e premoltiplicata per la trasposta del medesimo vettore (riga):

$$x'Ax$$

Si può senza perdita di generalità assumere che la matrice di una forma quadratica sia simmetrica, dato che il suo valore resta immutato qualora si sostituisca la matrice A con la matrice simmetrica $0.5(A + A')$.

$$0.5x'(A' + A)x = 0.5x'A'x + 0.5x'Ax = x'Ax$$

essendo lo scalare $x'A'x$ uguale alla sua trasposta $x'Ax$. Alcune forme quadratiche $x'Ax$ sono sempre positive per qualsiasi vettore $x \neq 0$. Un semplice esempio è costituito dalla forma quadratica con matrice diagonale i cui elementi diagonali sono tutti positivi:

$$x'Dx = \sum_{i=1}^n d_i x_i^2$$

dove il termine $d_i > 0$ indica l' i -esimo elemento diagonale di D . Le forme quadratiche $x'Ax$ sono denominate definite positive (e la corrispondente matrice A viene definita positiva) se si ha $x'Ax > 0$ per ogni $x \neq 0$ e semidefinite positive se $x'Ax \geq 0$ per ogni $x \neq 0$.

Esercizio.

Generare una matrice di numeri casuali di dimensione $n \times p$. Denominare questa matrice X . Assegnare ad n e p numeri interi a piacere (con $n > p$). Calcolare la matrice di covarianze S . Verificare empiricamente tramite 10000

simulazioni del vettore x (generato con qualsiasi combinazione di numeri) che la forma quadratica $x'Sx$ è definita positiva. Fare lo storing dei 10000 scalari $x'Sx$ in un vettore di dimensione 10000×1 denominato **formaquad**. Mostrare il valore più piccolo di $x'Sx$ delle 10000 simulazioni.

Soluzione

```
n=100;

p=3;

X=randn(n,p);

S=cov(X);

% nsimul = numero di simulazioni (ripetizioni dell'esperimento)
nsimul=10000;

% inizializzo il vettore formaquad che conterrà in
% posizione i il risultato di x'Sx per la simulazione i
formaquad=zeros(nsimul,1);

for i=1:nsimul

    x=randn(p,1);

    formaquad(i)=x'*S*x;

    assert(formaquad(i)>0,"Errore di programmazione la forma " + ...
        "quadratica x'Sx è definita positiva")

end

% Calcolo e mostro il minimo di formaquad
minfq=min(formaquad);

disp("Valore più piccolo di x'Sx in nsimul simulazioni")

disp(minfq)
```

Il criterio di Sylvester afferma che una matrice A simmetrica reale di dimensione $n \times n$ è definita positiva se

$$d_i > 0 \quad \text{per ogni } i \quad (1.6)$$

dove d_i è il determinante (minore) della matrice che si ottiene cancellando da A le ultime $n - i$ righe e le ultime $n - i$ colonne. Se vale la disuguaglianza debole $d_i \geq 0$ allora la matrice si dice semidefinita positiva. Similmente una matrice si dice definita negativa se per ogni i $(-1)^i d_i > 0$. Se per ogni i $(-1)^i d_i \geq 0$ la matrice si dice semidefinita negativa.

Esercizio: scrivere un codice che data una matrice A quadrata simmetrica determini, tramite il criterio di Sylvester, se questa matrice è (semi) positiva oppure (semi) definita negativa oppure indefinita. Fare scrivere nella command tramite l'istruzione `disp` il tipo di matrice. Soluzione

```
% Genero una matrice simmetrica
p=7;
A=randn(p,p);
A=A+A';

% p = numero di righe o colonne della matrice di input
p=size(A,2);

% Inizializzo il vettore riga che conterrà i minori principali
MinoriPrincipali=zeros(1,p);

% Trovo tutti i minori principali
seqp=1:p;
```

```

for i=seqp
    MinoriPrincipali(i)=det(A(1:i,1:i));
end

if all(MinoriPrincipali>0)
    disp("La matrice è definita positiva")
elseif all(MinoriPrincipali>=0)
    disp("La matrice è semidefinita positiva")
elseif all((-1).^seqp.*MinoriPrincipali>0)
    disp('La matrice è definita negativa')
elseif all((-1).^seqp.*MinoriPrincipali>=0)
    disp('La matrice è semidefinita negativa')
else
    disp('La matrice è indefinita')
end

```

Osservazione: si può dimostrare che se la matrice della forma quadratica è una matrice di varianze e covarianze oppure di correlazione, la forma quadratica associata è sempre semidefinita positiva.

Esempi di codice per generare matrici definite positive, semidefinite e definite negative e/o per chiedere all'utente di inserire in maniera interattiva una matrice è riportato nella Tabella 1.1.

Osservazione: se la matrice della forma quadratica è idempotente allora si parla di forma quadratica idempotente.

Esempio di generazione di una matrice A simmetrica definita positiva	<code>n=200; p=7; X=randn(n,p); A=X'*X;</code>
Esempio di generazione di una matrice A simmetrica definita negativa	<code>n=100; p=5; X=randn(n,p); A=-cov(X);</code>
Esempio di generazione di una matrice A simmetrica semidefinita positiva	<code>n=1000; p=6; X=[randn(n,p) zeros(n,1) randn(n,2)]; A=cov(X);</code>
Esempio di codice per chiedere all'utente di inserire una determinata matrice	<code>A=input("Inserire la matrice da testare");</code>

Tabella 1.1: Esempi di generazione di matrici definite positive, semidefinite e definite negative.

La forma quadratica $x'Ax$ può essere scritta come:

$$x'Ax = a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + \cdots + a_{nn}x_{nn}^2 = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_ix_j$$

Questo non è altro che un polinomio omogeneo di grado 2 (il totale delle potenze di ciascun termine è esattamente uguale a 2). Questo polinomio fa parte della famiglia delle coniche⁴

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0.$$

⁴Le coniche sono curve piane che si ottengono dall'intersezione tra un piano ed un cono a due falde. Si veda ad esempio https://it.wikipedia.org/wiki/Rappresentazione_matriciale_delle_coniche

La conica si riduce ad un'ellisse se

$$\det \begin{pmatrix} a & b \\ b & c \end{pmatrix} = ac - b^2 > 0 \quad d = 0 \quad e = 0 \quad f < 0$$

Quindi, nel caso in cui la matrice della forma quadratica sia definita positiva (come ad esempio nel caso della matrice di covarianze S), l'equazione $x'Sx = c^2$ definisce sempre un'ellisse in due dimensioni e un ellissoide in presenza di più di due dimensioni che racchiude un determinato numero di punti⁵.

La distribuzione normale p -variata presenta la seguente densità:

$$\begin{aligned} f(x, \mu, \Sigma) &= \frac{1}{2\pi \det(\Sigma)^{0.5}} \exp \left(-\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right) \\ &= \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{0.5}} \exp -0.5Q \end{aligned}$$

$x = (x_1, \dots, x_p)'$, $\mu = E(x)$ e $\Sigma = \text{var}(x)$. La forma quadratica $Q = (x - \mu)' \Sigma^{-1} (x - \mu)$ che compare nell'argomento dell'esponenziale è una quantità non negativa. I valori di x e y per i quali la densità di probabilità è costante sono quelli per cui è costante Q . Questi punti definiscono un ellissoide nello spazio a p dimensioni ed un'ellisse quando $p = 2$ e $x = (x_1, x_2)'$.

Esercizio

⁵Nell'equazione

$$x'Sx = c^2.$$

abbiamo indicato il termine noto con c^2 invece che con c per stressare il fatto che questo è un termine maggiore di zero (in quanto la forma quadratica è definita positiva).

Dato il vettore delle medie $\mu = (2 \ 3)'$ e la matrice di covarianze

$$\Sigma = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{pmatrix}$$

mostrare le isolinee o curve di livello⁶ della funzione $Q(x) = Q(x_1, x_2)$ in corrispondenza delle altezze $Q(x) = 0, 20, \dots, 200$, aggiungendo il relativo testo.

$$Q(x) = (x - \mu)' \Sigma^{-1} (x - \mu)$$

Aggiungere al grafico la barra di colore⁷. Considerare per x_1 le coordinate -15, -14.95, ..., 15 e per x_2 le coordinate -13, -12.95, ..., 18.

Soluzione

```
mu = [2 3];
Sigma=[0.25 0.3; 0.3 1];
% Precalcolo l'inversa
SigmaInv = inv(Sigma);

x1 = -15:.05:15;
x2 = -13:.05:18;

l1=length(x1);
```

⁶Le isolinee o curve di livello sono il luogo dei punti in cui la funzione presenta lo stesso valore.

⁷ Nei grafici 3D è spesso utile affiancare al grafico una barra di colore che indica la corrispondenza tra colori e valori di altezza z , usando la funzione **colorbar**:

- **colorbar** oppure **colorbar('vert')** inserisce la colorbar in verticale
- **colorbar('horiz')** inserisce la colorbar in orizzontale

```
l2=length(x2);  
% Si inizializza la matrice Q che conterrà  
% il valore della funzione in corrispondenza di ogni  
% combinazione di x1 e x2  
Q=zeros(l1,l2);  
for i=1:l1  
    for j=1:l2  
        x=[x1(i) x2(j)]-mu;  
        Q(i,j)=x*SigmaInv*x';  
    end  
end  
% opzione 'ShowText','on' per mostrare il testo  
contour(x1,x2,Q',0:20:200,'ShowText','on')  
xlabel('x1')  
ylabel('x2')  
% Aggiunta della barra di colore  
colorbar
```

Il codice di cui sopra produce il grafico riportato nella Figura 1.4.

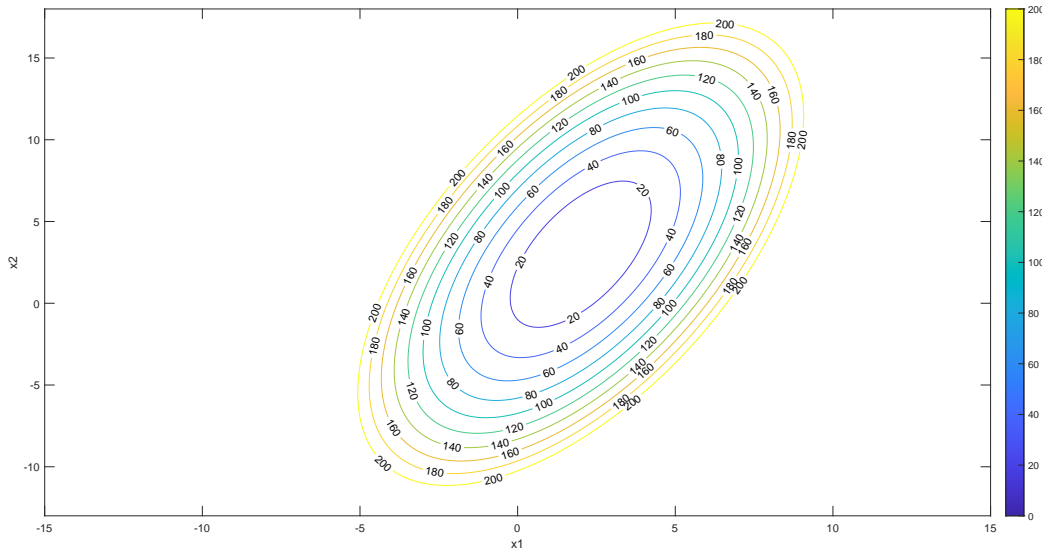


Figura 1.4: Curve di livello della funzione biviariata $Q(x) = (x - \mu)' \Sigma^{-1} (x - \mu)$, $x = (x_1, x_2)'$ ottenute tramite la funzione `contour`. Questa funzione proietta i punti di uguale altezza in 3D (curve di livello) su un piano 2D sottostante.

1.1.8 Estrazione degli elementi tramite forme quadratiche

Sia data una matrice A rettangolare di dimensione $n \times p$, l'elemento i, j di A , a_{ij} , può essere scritto in maniera matriciale come:

$$e_i' A e_j = a_{ij} \quad (1.7)$$

dove e_i e e_j sono vettori colonna elementari di lunghezza rispettivamente pari a n e p .

Esercizio: estrazione elementi in maniera matriciale

Generare una matrice di numeri casuali di dimensione $n \times p$, denominata A , Premoltiplicare e postmoltiplicare la matrice precedente in modo appropriato per poter estrarre l'elemento che si trova all'incrocio della riga i e della

colonna j (porre $i=2$, $j=5$, $n = 10$ e $p = 6$)

Soluzione

```
n=10;
p=6;
A=randn(n,p);

% Definizione variabili i,j
i=2;
j=5;
% epre = vettore elementare e_i di lunghezza n (tutti elementi uguali a
% zero tranne quello in posizione i che risulta uguale ad 1)
epre=zeros(n,1);
epre(i)=1;
% epost = vettore elementare e_j di lunghezza p (tutti elementi uguali a
% zero tranne quello in posizione j che risulta uguale ad 1.
epost=zeros(p,1);
epost(j)=1;

disp(['Estrazione elemento(' num2str(i) ',' num2str(j) ') ' ...
      'di A in maniera matriciale'])
disp(epre'*A*epost)
disp(['Estrazione elemento(' num2str(i) ',' num2str(j) ') ' ...
      'di A tramite l''istruzione A(i,j)'])
disp(A(i,j))
```

1.1.9 Matrici ortogonali

Una matrice quadrata di dimensione V si dice ortogonale se la sua trasposta è uguale alla sua inversa: $V' = V^{-1}$. Il determinante di una matrice ortogonale⁸ è 1 oppure -1.

L'istruzione `gallery('orthog',n,k)` dove n determina l'ordine della matrice quadrata e k è un numero naturale compreso tra 1 e 6 permette di avere una serie di matrici ortogonali.

Esercizi sulle matrici ortogonali.

Tramite la funzione `gallery` generare una matrice ortogonale di ordine 5. Per ogni valore intero del secondo argomento k della funzione `gallery` compreso tra 1 e 6, verificare che:

1. la matrice sia ortogonale;
2. il suo determinante sia uguale a 1 oppure a -1

Soluzione

```
n=5;
for j=1:6
    A = gallery('orthog',n,j);
```

⁸Questo si può dimostrare come segue:

$$1 = \det(I) = \det(G \cdot G') = \det(G)\det(G') = (\det(G))^2$$

Gli unici due numeri il cui quadrato è 1 sono -1 e 1 .

```

% Verifico l'ortogonalità di A per ogni valore di k
maxdiff=max(abs(A'*A-eye(n)),[],'all');
assert(maxdiff<1e-12,"Errore di programmazione nella verifica " + ...
    "dell'ortogonalità della matrice generata tramite gallery" + ...
    "o valore di k diverso da 1, 2, ..., 6")
maxdiff=max(abs(A*A'-eye(n)),[],'all');
assert(maxdiff<1e-12,"Errore di programmazione nella verifica " + ...
    "dell'ortogonalità della matrice generata tramite gallery" + ...
    "o valore di k diverso da 1, 2, ..., 6")
detA=det(A);
% Controllo che il determinante di A sia 1 oppure -1
assert(min(abs(detA-[-1 1]))<1e-12,"Errore di programmazione " + ...
    "Il determinante di una matrice ortogonale è " + ...
    "uguale a 1 oppure uguale a -1")
end

```

1.1.10 Moltiplicazione di matrici trasposte

Date 3 matrici conformabili A , B e C , si ha la seguente regola.

$$(ABC)' = C'B'A' \quad (1.8)$$

Esercizio: Date 3 matrici A , B e C di dimensione $m \times n$, $n \times o$, $o \times p$, verificare la precedente uguaglianza. Generare i numeri m, n, o, p utilizzando la funzione `randi` inserendo il vincolo $\max(m, n, o, p) \leq 9$. Per replicabilità dei risultati utilizzare il seed 12345.

Soluzione

```
rng(12345);  
% Estraggo numeri interi positivi nell'intervallo 1-9  
m=randi(9,1);  
% Alternativamente si poteva utilizzare l'istruzione m=randi(9);  
n=randi(9,1);  
o=randi(9,1);  
p=randi(9,1);  
% Creo le matrici richieste utilizzando distribuzioni a piacere  
% A = mxn numeri generati dalla distribuzione N(0,1)  
A=normrnd(0,1,m,n);  
% B = nxo numeri generati dalla distribuzione U(1,3)  
B=unifrnd(1,3,n,o);  
% C= oxp numeri generati dalla distribuzione Chi2  
% con 6 gradi di libert   
C=chi2rnd(6,o,p);  
D=(A*B*C)';  
Dchk=(C'*B'*A');  
disp(' Verifica che (A*B*C)''-(C''*B''*A'') = zeros(p,m)')  
disp(D-Dchk)  
maxdiff=max(abs(D-Dchk),[],'all')<1e-10;  
assert(maxdiff,'Errore di programmazione nella verifica ...' + ...  
        'della regola della moltiplicazione della matrice trasposta')
```

Per quanto riguarda il determinante si ha che

$$\det(ABC) = |ABC| = |A||B||C|$$

1.1.11 Moltiplicazione di matrici inverse

Nel caso della matrice inversa, se A , B e C sono matrici quadrate invertibili⁹ della stessa dimensione

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$$

Osservazione: Matlab fornisce due tecniche per calcolare l'inversa di una matrice: la funzione `inv` (ad esempio `inv(A)`) oppure l'elevazione della matrice alla potenza -1 (ad esempio A^{-1}).

Esercizio

Date tre matrici quadrate di ordine n contenenti numeri casuali estratti da qualsiasi distribuzione, verificare che $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$. Soluzione

`n=8;`

`% Di seguito senza perdita di generalità si utilizzano`

`% numeri estratti da N(0,1)`

`A=randn(n,n);`

`B=randn(n,n);`

`C=randn(n,n);`

`invABC=inv(A*B*C);`

⁹Una matrice ammette l'inversa se il suo determinante è diverso da zero.


```

invCBA=inv(C)*inv(B)*inv(A);
maxdiff=max(abs(invABC-invCBA),[],"all")<1e-10;
assert(maxdiff,"Errore di programmazione nella verifica " + ...
        "della regola della moltiplicazione di matrici inverse")

```

1.1.12 La trasposta dell'inversa

Se una matrice quadrata A è invertibile, allora anche A' è invertibile. Inoltre, si ha che:

$$(A')^{-1} = (A^{-1})'$$

Questo risultato discende immediatamente dalla regola di trasposizione del prodotto. Applicando la regola nell'equazione (1.8), si ha:

$$A'(A^{-1})' = (A^{-1}A)' = I' = I$$

Quindi, se $A'(A^{-1})' = I$ significa che $(A^{-1})' = (A')^{-1}$

Esercizio.

Generare una matrice quadrata di dimensione $n \times n$ e verificare l'identità di cui sopra.

Soluzione

Soluzione

```

n=10;
A=randn(n,n);
Atraspinv=(A')^-1;
Ainvtrasp=(A^-1)';

```

```

maxdiff=max(abs(Atraspinv(:)-Ainvtrasp(:)));
assert(maxdiff<1e-12,"Errore di programmazione nella verifica " + ...
      "delle proprietà dell'inversa della trasposta")

```

1.1.13 La traccia

Data una matrice A quadrata, $tr(A)$ è la somma degli elementi sulla diagonale principale. Data una matrice A di dimensione $m \times n$ ed una matrice B di dimensione $n \times m$ si ha

$$tr(AB) = tr(BA)$$

L'operatore traccia è invariante rispetto ad una permutazione ciclica nel senso che

$$tr(ABCD) = tr(BCDA) = tr(CDAB) = tr(DABC).$$

Date due matrici X e Y di dimensione $n \times p$ la traccia di un prodotto può essere scritta come segue:

$$tr(X'Y) = tr(XY') = tr(Y'X) = tr(YX') = \sum_{i=1}^n \sum_{j=1}^p x_{ij}y_{ij} \quad (1.9)$$

Esercizio.

Generare due matrici di dimensioni $n \times p$ e verificare le identità presenti nell'equazione (1.9).

Soluzione

```
p=5;
```

```
X=randn(n,p);
```

```

Y=randn(n,p);
% i 5 elementi del vettore tracce contengono i diversi modi
% di implementazione della traccia del prodotto delle due matrici
tracce=zeros(5,1);
tracce(1)=trace(X'*Y);
tracce(2)=trace(X*Y');
tracce(3)=trace(Y'*X);
tracce(4)=trace(Y*X');
tracce(5)=sum(X.*Y,'all');

maxdiff=max(tracce)-min(tracce);
assert(maxdiff<1e-12,"Errore di programmazione nella verifica " + ...
    "delle proprietà della traccia")

```

Si noti che se si pone $X = Y$ dall'equazione (1.9) si ottiene che la somma dei quadrati degli elementi di una generica matrice X si può scrivere come:

$$tr(X'X) = tr(XX') = \sum_{i=1}^n \sum_{j=1}^p x_{ij}^2 \quad (1.10)$$

Esercizio

Verificare che la somma dei quadrati degli elementi di una generica matrice X di dimensione $n \times p$ è uguale alla traccia di $X'X$ oppure alla traccia di XX' . Per replicabilità dei risultati utilizzare il seed 4567, $n = 10$ e $p = 3$.

Soluzione

```

rng(4567)
n=10;

```

```

p=3;
X=randn(n,p);
trXtraspostoX=trace(X'*X);
trXXtrasposto=trace(X*X');
sommaqua=sum(X.^2,"all");
maxdiff=max([abs(trXtraspostoX-trXXtrasposto),...
    abs(trXtraspostoX-sommaqua)])<1e-11;
nomitable=["tr(X'X)" "tr(XX')" "Somma calcolata direttamente"];
St=array2table([trXtraspostoX trXXtrasposto sommaqua],...
    'VariableNames',nomitable);

assert(maxdiff,"Errore di programmazione nella verifica " + ...
    "del calcolo della somma dei quadrati " + ...
    "tramite tr(X'X) oppure tr(X*X')")

```

Osservazione: la radice quadrata della somma dei quadrati degli elementi di una matrice A : ossia $\sqrt{\text{tr}(A'A)}$ è nota in letteratura come norma di Frobenius. L'istruzione

```
norm(A,'fro')
```

consente di calcolare immediatamente questo numero.

1.1.14 Dipendenza, indipendenza lineare e base di uno spazio vettoriale

Definizione di combinazione lineare: dati p vettori ad n componenti: x_1, x_2, \dots, x_p , si dice combinazione lineare dei vettori con coefficienti c_i la seguente

espressione:

$$c_1x_1 + c_2x_2 + \dots + c_px_p$$

Se i vettori sono inseriti in una matrice $X = (x_1, \dots, x_p)$ di dimensione $n \times p$ e si definisce il vettore $c' = (c_1, c_2, \dots, c_p)$, allora la combinazione lineare si scrive in forma matriciale come Xc .

Definizione di dipendenza lineare di un vettore da altri vettori.

Sia y un vettore ad n componenti. Se y è esprimibile come combinazione lineare di p vettori (sempre ad n componenti) x_1, x_2, \dots, x_p , si dice che y è linearmente dipendente dai p vettori considerati. Formalmente, y è linearmente dipendente da x_1, x_2, \dots, x_p , se vale per qualche insieme di coefficienti c_i :

$$y = c_1x_1 + c_2x_2 + \dots + c_px_p = Xc.$$

Viceversa, se non esiste un insieme di coefficienti c_i che consentono di esprimere il vettore y come combinazione lineare dei vettori x_i , allora y si dice linearmente indipendente dall'insieme di vettori considerati.

Si definisce base di uno spazio vettoriale l'insieme dei vettori linearmente indipendenti che generano lo spazio. In modo equivalente, ogni elemento dello spazio vettoriale può essere scritto in modo unico come combinazione lineare dei vettori appartenenti alla base. Più formalmente, sia V uno spazio vettoriale, l'insieme degli elementi di V , v_1, v_2, \dots, v_n è una base di V :

- se i vettori v_1, v_2, \dots, v_n sono linearmente indipendenti e la relazione

$$\sum_{i=1}^n a_i v_i = a_1 v_1 + a_2 v_2 + \dots + a_n v_n = 0$$

è verificata solo se i numeri a_1, a_2, \dots, a_n sono tutti uguali a zero.

- i vettori v_1, v_2, \dots, v_n generano V , ovvero:

$$V := \{a_1 v_1 + \dots + a_n v_n \mid a_1, \dots, a_n \in R^n\}$$

il simbolo ‘:=’ si deve leggere: “è uguale per definizione”

I numeri a_1, a_2, \dots, a_n (coefficienti della combinazione lineare) sono le coordinate rispetto alla base scelta. Sia R^n un campo. Si definisce base canonica di R^n l'insieme di vettori:

$$e_1 = (1, 0, \dots, 0)' \quad e_2 = (0, 1, \dots, 0)' \quad \dots \quad ; e_n = (0, 0, \dots, 1)'$$

Ogni vettore in $w \in R^n$ si può scrivere come combinazione lineare dei vettori della base canonica:

$$\sum_{i=1}^n a_i e_i$$

Il vettore: $a = (a_1, a_2, \dots, a_n)'$ è il vettore delle coordinate di w rispetto alla base canonica.

Ad esempio, i vettori $e_1 = (1, 0)'$ e $e_2 = (0, 1)'$ sono una base di R^2 , infatti ogni vettore $c = (a, b)'$ si scrive come:

$$c = \begin{pmatrix} a \\ b \end{pmatrix} = a \begin{pmatrix} 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 0 \\ 1 \end{pmatrix} = a e_1 + b e_2$$

Ogni spazio vettoriale reale contiene infinite basi. Ogni vettore dello

spazio vettoriale può essere rappresentato in modo diverso a seconda della base che si sceglie. In termini matriciali, se x_1, x_2, \dots, x_n sono n vettori nello spazio R^p nella base data dai vettori v_1, v_2, \dots, v_p , significa che ognuno di questi vettori può essere scritto come

$$x_i = y_{i1}v_1 + y_{i2}v_2 + \dots + y_{ip}v_p = (v_1 \ v_2, \dots, v_p) \begin{pmatrix} y_{i1} \\ \dots \\ y_{ip} \end{pmatrix} = V y_i \quad (1.11)$$

oppure che

$$x'_i = y'_i V'$$

Se i vettori x'_1, \dots, x'_n , e y'_1, \dots, y'_n formano rispettivamente le n righe della matrice X e della matrice Y abbiamo che:

$$X = \begin{pmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{pmatrix} = \begin{pmatrix} y'_1 \\ y'_2 \\ \dots \\ y'_n \end{pmatrix} V' = Y V' \quad (1.12)$$

La scrittura $Y V'$ significa, quindi, che le righe della matrice X sono espresse in funzione della base v_1, \dots, v_p e che la matrice Y contiene le coordinate di ogni vettore x_1, \dots, x_n nella base V .

Nell'esercizio che segue si mostra come rappresentare un insieme di vettori in due basi diverse. Per ulteriori approfondimenti si veda la pagina https://en.wikipedia.org/wiki/Change_of_basis.

Esercizio

La matrice X che segue contiene (nelle righe) 4 vettori nello spazio R^2 .

$X = [1.7 \ 3.5; \ 1 \ 5; \ -2 \ 2.5; \ 5 \ 1]$;

Calcolare le nuove coordinate nel sistema ortogonale determinato dai vettori v_1 e v_2

$$v_1 = \begin{pmatrix} 0.5156 \\ 0.8569 \end{pmatrix} \quad v_2 = \begin{pmatrix} -0.8569 \\ 0.5156 \end{pmatrix}$$

Osservazione: $V = (v_1 \ v_2)$ è una matrice ortogonale, ossia $V'V = I_2$.

Suddividere la finestra grafica in 4 pannelli. Nel pannello in alto a sinistra rappresentare i vettori nella base canonica. Aggiungere ai punti con il colore rosso le etichette A, B, e le rispettive coordinate nella base canonica. Aggiungere alla rappresentazione grafica i vettori della base canonica moltiplicati per 5 con il colore rosso (v. pannello in alto a sinistra della Figura 1.5).

Nel pannello in alto a destra rappresentare i vettori della nuova base con colore nero ed aggiungere nelle etichette dei punti le coordinate nella nuova base con il colore nero (v. pannello in alto a destra della Figura 1.5).

Nel pannello in basso a sinistra rappresentare i punti nel nuovo sistema di assi cartesiani v_1 e v_2 . Aggiungere nelle etichette dei punti le coordinate nella nuova base con il colore nero. Aggiungere gli assi cartesiani v_1 e v_2 (v. pannello in basso a sinistra della Figura 1.5).

Soluzione

```
X=[1.7 3.5; 1 5; -2 2.5; 5 1];
```

```
[n,p]=size(X);
```

```
zern=zeros(n,1);
```

```
zerp=zeros(p,1);
```

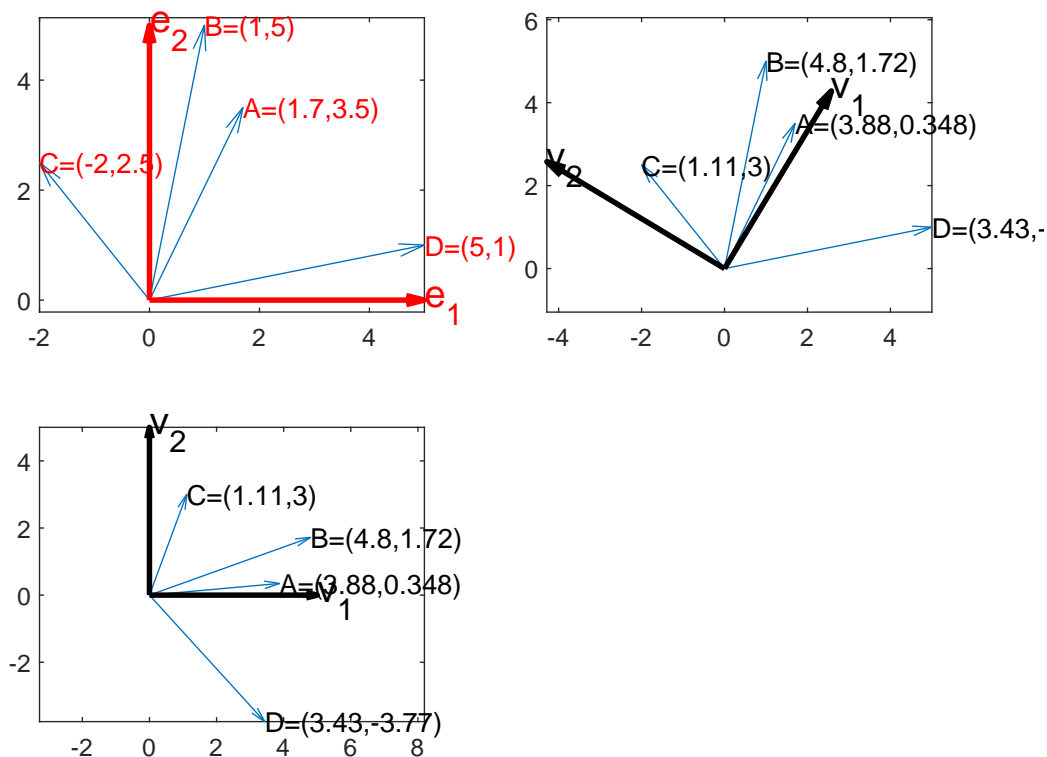



Figura 1.5: Dettagli relativi al cambiamento di base di un insieme di vettori \mathbb{R}^2 . Pannello in alto a sinistra: coordinate dei punti nello spazio originale (base canonica e_1 e_2). Pannello in alto a destra: rappresentazione dei nuovi assi cartesiani v_1 v_2 . I numeri dentro le parentesi sono le coordinate dei punti nel nuovo sistema di riferimento. Pannello in basso a sinistra: rappresentazione dei punti nel nuovo sistema di assi cartesiani v_1 , v_2 .

```
% Etichette dei punti A, B, C, ....
labx=char((( 'A'+0):( 'A'+n-1))');

% lwd = Linewidth delle frecce associate agli assi
lwd=2;

% colororiginalaxis = colore degli assi originali
colororiginalaxis='r';

% colnewaxis = colore dei nuovi assi cartesiani
```

```

colnewaxis='k';

% fsize = FontSize delle label degli assi
fsize=14;

% kk = scalare che determina la lunghezza
% degli assi cartesiani
kk=5;

% base canonica (matrice identità)
E=eye(p);

nr=2;
nc=2;

%% Pannello in altro a sinistra: punti nello spazio originale
subplot(nr,nc,1)

% quiver(zern,zern,X(:,1),X(:,2)) traccia le frecce da 0,0 ai punti X(:,1) e
% X(:,2). Il terzo argomento di quiver settato su 'off' serve per fare in
% modo che la lunghezza delle frecce non venga riscalata in automatico.
quiver(zern,zern,X(:,1),X(:,2),'off')
textOrigCoo=labx+ "(" + string(num2str(X(:,1),3)) + ...
    ", " + string(num2str(X(:,2),3)) + ")";
% Gli spazi vuoti dentro textOrigCoo vengono eliminati
textOrigCoo=strrep(textOrigCoo," ","");
text(X(:,1),X(:,2),textOrigCoo,'HorizontalAlignment',...
    'left','Color',colororiginalaxis)

```

```

hold('on')

% Assi cartesiani base canonica
quiver(zerp,zerp,kk*E(:,1),kk*E(:,2),'off',...
       'LineWidth',lwd,'Color',coloriginalaxis)

% label degli assi della base canonica
labaxis=["e_1";"e_2"];
text(kk*E(:,1),kk*E(:,2),labaxis,...
     'HorizontalAlignment','left','Color',...
     coloriginalaxis,'FontSize',fsize)

% scala uguale nei due assi
axis equal

%% Pannello in altro a destra: vettori nella nuova base
% e coordinate dei punti nella nuova base
subplot(nr,nc,2)
quiver(zern,zern,X(:,1),X(:,2),'off')
hold('on')
V=[ 0.5156   -0.8569; 0.8569    0.5156];
Vtra=V';
% Nuovi assi cartesiani ortogonali v_1 v_2
quiver(zerp,zerp,kk*Vtra(:,1),kk*Vtra(:,2), ...
       'off','LineWidth',lwd,'Color',colnewaxis)
% label degli assi
labnewaxis=["v_1";"v_2"];

```

```

text(kk*Vtra(:,1),kk*Vtra(:,2),labnewaxis,...
     'HorizontalAlignment','left', ...
     'Color',colnewaxis,'FontSize',fsize)

% Coordinate dei punti nel nuovo riferimento
% ortogonale V(:,1), V(:,2)
Y=X*V;
textNewCoo=labx+ "(" + string(num2str(Y(:,1),3)) + ...
            "," + string(num2str(Y(:,2),3))+")";
textNewCoo=strrep(textNewCoo," ","");
% L'istruzione di seguito inserisce l'indicazione
% delle coordinate dei punti nel nuovo sistema di
% riferimento (le etichette vengono mostrate in
% corrispondenza di X(:,1) e X(:,2) ossia delle
% vecchie coordinate
text(X(:,1),X(:,2),textNewCoo,'Color',colnewaxis)
axis equal

%% Pannello in basso a sinistra: vettori nella nuova base
subplot(nr,nc,3)
quiver(zern,zern,Y(:,1),Y(:,2),'off')
hold('on')
quiver(zerp,zerp,kk*E(:,1),kk*E(:,2),...
      'off','LineWidth',lwd,'Color',colnewaxis)

```

```

text(Y(:,1),Y(:,2),textNewCoo,'Color',colnewaxis)
text(kk*E(:,1),kk*E(:,2),labnewaxis,...
      'HorizontalAlignment','left','Color',...
      colnewaxis,'FontSize',fsize)

```

axis equal

1.1.15 Il rango

Il rango di una matrice $\text{rank}(A)$ di dimensione $n \times p$ è il massimo numero di righe (o colonne) linearmente indipendenti. Il rango gode della seguente proprietà:

- Il rango di una matrice A di dimensione $m \times n$ è minore o uguale al minimo tra m e n . Ad esempio, il rango di un vettore ad n componenti è 1.
- $\text{rank}(A) = \text{rank}(A'A) = \text{rank}(AA')$;
- $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$
- $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$ quando A e B sono della stessa dimensione. Questa proprietà significa che una matrice di rango k può essere espressa come somma di k matrici di rango 1, ma non di un numero inferiore.

- Se A è una matrice di dimensione quadrata di dimensione n , il rango di A è pari ad n se il determinante di A è diverso da zero.

Osservazione: la funzione per calcolare il rango di una matrice si chiama **rank**.

Esercizi sul rango.

Generare 4 vettori di numeri casuali x_1, x_2, x_3, x_4 , di dimensione $n \times 1$ (con $n \geq 4$) e verificare che

- il rango della matrice $A = x_1 x_1'$ di dimensione $n \times n$ è 1

Osservazione: $\text{rank}(A) = \text{rank}(x_1 x_1') = \text{rank}(\sum_{i=1}^n x_{i1}^2) = 1$

- il rango della matrice $A = x_1 x_2'$ di dimensione $n \times n$ è 1

Osservazione: $\text{rank}(A) = \text{rank}(x_1 x_2') \leq \min(\text{rank}(x_1), \text{rank}(x_2)) = \min(1, 1) = 1$

- il rango della matrice $A = x_1 x_1' + x_2 x_2'$ di dimensione $n \times n$ è 2

- il rango della matrice $A = \sum_{i=1}^4 x_i x_i'$ di dimensione $n \times n$ è quattro

- il rango della matrice $A = \sum_{i=1}^4 x_3 x_3'$ di dimensione $n \times n$ è uno

Soluzione

`n=7;`

`x1=randn(n,1);`

`x2=randn(n,1);`

```

x3=randn(n,1);
x4=randn(n,1);
% rango della matrice A =x1 x1'
A=x1*x1';
rankA=rank(A);
assert(rankA==1,"Errore di programmazione rango A=x1*x1' è 1")
% rango della matrice A =x1 x2'
B=x1*x2';
rankB=rank(B);
assert(rankB==1,"Errore di programmazione rango B=x2*x2' è 1")
% rango della matrice C =x1 x1'+ x2 x2'
C=x1*x1'+x2*x2';
rankC=rank(C);
assert(rankC==2,"Errore di programmazione rango x1 x1'+ x2 x2' è 2")
% rango della matrice D =x1*x1'+x2*x2'+x3*x3'+x4*x4'
D=x1*x1'+x2*x2'+x3*x3'+x4*x4';
rankD=rank(D);
assert(rankD==4,"Errore di programmazione rango x1 x1'+ " + ...
    " x2 x2' +x3*x3'+x4*x4' è 4")
% il rango della matrice E =\sum_{i=1}^4 x_3 x_3' è uno
E=4*(x3*x3');
rankE=rank(E);
assert(rankE==1,"Errore di programmazione rango x3 x3'+ " + ...
    " x3 x3' +x3*x3'+x3*x3' è 1")

```

1.1.16 Sistemi di equazioni lineari

I sistemi lineari sono gruppi di equazioni in cui ciascuna di esse è un'equazione di primo grado in una o più incognite. Un sistema lineare può essere rappresentato nel calcolo matriciale nella forma $Ax = b$, dove A è la matrice $m \times n$ dei coefficienti, x è il vettore delle variabili incognite (x_1, \dots, x_n) e b è il vettore dei termini noti. Si dice soluzione del sistema ogni vettore (x_1, \dots, x_n) le cui coordinate soddisfano simultaneamente tutte le equazioni del sistema. Il sistema lineare $Ax = b$ ammette una sola soluzione se la matrice A è non singolare (ossia se il suo determinante è diverso da zero). In tal caso $x = A^{-1}b$. Si parla di sistema lineare omogeneo quando il vettore dei termini noti è composto soltanto da n zeri. Questa caratteristica, all'apparenza poco rilevante, ha una grande ripercussione sull'insieme delle soluzioni del sistema nel senso che un sistema lineare omogeneo ammette sempre almeno una soluzione data dal caso in cui tutti gli elementi di x sono pari a 0.

Se la matrice A è quadrata e non singolare di ordine n , il sistema avrà un'unica soluzione $x = A^{-1}0_{n \times 1} = 0_{n \times 1}$, ovvero la soluzione banale. Affinché esistano anche soluzioni non banali dovrà essere $\det(A) = 0$.

Esercizio:

Dato il seguente sistema di equazioni lineari:

$$\begin{cases} 3x_1 + 5x_2 + 4x_3 &= 25 \\ 2x_1 + 10x_2 + x_3 &= 25 \\ 2x_1 + 2x_2 + 9x_3 &= 33 \end{cases}$$

inserirlo in maniera matriciale nella forma $Ax = b$. Verificare che il rango di A sia uguale a 3 (ossia che il sistema ammetta una sola soluzione). Calcolare $x = (x_1, x_2, x_3)'$.

Soluzione

```
% A =matrice dei coefficienti
A=[ 3 5 4; 2 10 1; 2 2 9];
% b = vettore contenente i termini noti del sistema
b=[25; 25; 33];
% Verifica del rango
assert(rank(A)==3,['La matrice non è di rango pieno' ...
    'Il sistema ammette infinite soluzioni'])
% risoluzione del sistema
x=inv(A)*b;
format long
disp("Soluzione del sistema di equazioni lineari " + ...
    "tramite istruzione inv(A)*b")
disp(x)
% Osservazione: computazionalmente il modo più efficiente per risolvere il
% sistema equazioni lineari è tramite l'operatore \
x=A\b;
disp("Soluzione del sistema di equazioni lineari " + ...
    "tramite istruzione A\b")
disp(x)
```

Si noti che il primo `disp(x)` produce

Soluzione del sistema di equazioni lineari tramite istruzione `inv(A)*b`

1.0000000000000004

2.0000000000000000

3.0000000000000000

A contrario il secondo `disp(x)` derivante dall'istruzione `x=A\b` produce

Soluzione del sistema di equazioni lineari tramite istruzione `A\b`

1

2

3

1.2 Autovalori e autovettori

Data una generica matrice A di dimensione $m \times m$ e x un vettore colonna di lunghezza m , dato uno scalare λ (λ appartiene allo spazio dei numeri complessi) se si ha (per $x \neq 0$) che

$$Ax = \lambda x, \quad (1.13)$$

λ è detto autovalore ed il corrispondente vettore x è detto autovettore della matrice A associato all'autovalore λ . L'autovettore, quindi, è un vettore non nullo che, moltiplicato per una matrice A , diventa un multiplo di sé stesso.

Per una illustrazione grafica di questo concetto si veda il sito

<https://blogs.mathworks.com/cleve/2013/07/08/eigshow-week-1/>

Osservazione: vedremo nel seguito che in generale per la matrice A di dimensione $m \times m$ ci sono al massimo m valori di λ che soddisfano l'equazione (1.13).

Osservazione: gli autovettori di una matrice non sono unici: se x è un autovettore di A associato a λ anche cx , con c scalare qualsiasi è un autovettore di A associato a λ . $A(cx) = cAx = c\lambda x = \lambda(cx)$. Per fare in modo che l'autovettore sia definito unicamente, si impone che la somma dei quadrati degli elementi di x sia uguale ad 1. In simboli: $x'x = 1$ (norma unitaria, v. sezione 1.1.5).

1.2.1 Polinomio caratteristico

L'obiettivo di questa sezione è capire come si calcolano in pratica gli autovalori ed i relativi autovettori. Dall'espressione $Ax = \lambda x$, si ricava $Ax - \lambda x = (A - \lambda I_n)x = 0$, essendo I_n la matrice identità. Affinché esista $x \neq 0$ che soddisfa questa relazione, la matrice $A - \lambda I$ deve essere singolare cioè il suo determinante deve essere zero (v. sezione 1.1.16).

$$\det(A - \lambda I) = 0 \quad (1.14)$$

Gli autovalori di una matrice A sono le radici del polinomio caratteristico $p(\lambda)$ definito da

$$p(\lambda) = \det(A - \lambda I)$$

Ad esempio, se

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix}$$

$$A - \lambda I = \begin{pmatrix} 2 - \lambda & 2 \\ 1 & 3 - \lambda \end{pmatrix}$$

il polinomio caratteristico è:

$$p(\lambda) = |A - \lambda I| = (2 - \lambda)(3 - \lambda) - 2 = \lambda^2 - 5\lambda + 4 \quad (1.15)$$

Risolvendo l'equazione di secondo grado si ottiene:

$$\lambda_1 = 4 \quad \lambda_2 = 1$$

Per ogni autovalore trovato occorre risolvere il sistema lineare omogeneo

$$(A - \lambda I)x = 0$$

per calcolare l'autovettore corrispondente. Sostituendo il valore di λ_1 nell'equazione (1.13) si ottiene

$$\begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 4 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{cases} 2x_1 + 2x_2 = 4x_1 \\ 1x_1 + 3x_2 = 4x_2 \end{cases}$$

$$\begin{cases} 2x_2 = 2x_1 \\ x_1 = x_2 \end{cases}$$

Ovviamente il sistema è indeterminato e presenta infinite soluzioni. Per fare in modo che l'autovettore sia univocamente determinato si impone il vincolo che la somma dei quadrati dei suoi elementi sia uguale ad 1. In altri termini, si impone il vincolo di norma (v. equazione (1.1)) unitaria. Ponendo quindi a sistema:

$$\begin{cases} x_1 = x_2 \\ \|x\|^2 = x_1^2 + x_2^2 = 1 \end{cases}$$

si ottiene che $x_1 = x_2 = 1/\sqrt{2}$. Il primo autovettore (che definiamo v_1) corrispondente all'autovalore $\lambda_1 = 4$ è dato da

$$v_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

Con passaggi analoghi si ricava che il secondo autovettore (associato a $\lambda_2 = 1$) è dato dal vettore

$$v_2 = \begin{pmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix}$$

Osservazione: gli autovettori sono definiti a meno del segno nel senso che se x è un autovettore di A anche $-x$ è un autovettore di A in quanto

l'equazione

$$Ax = \lambda x$$

è verificata anche da

$$A(-x) = \lambda(-x) \quad (1.16)$$

Osservazione 1: quando A è di ordine $n \times n$, la sua equazione caratteristica diventa un polinomio di grado n -esimo nella variabile λ che, quindi, ammette n soluzioni $\lambda_1, \dots, \lambda_n$ reali o immaginarie a cui corrispondono n autovettori x_1, \dots, x_n . Anche nel caso in cui la matrice A sia costituita da elementi reali, gli autovalori non sono necessariamente reali. Tuttavia, si può dimostrare che se la matrice A è simmetrica le radici caratteristiche sono sempre reali ed il numero di autovalori non nulli coincide con il rango di A .

Osservazione 2: se la matrice A è simmetrica e di ordine p , gli autovettori v_i e v_j relativi a due autovalori distinti λ_i e λ_j sono a due a due ortogonali ossia $v_i'v_j = 0$ $i \neq j = 1, 2, \dots, p$. La matrice degli autovettori $V = (v_1, v_2, \dots, v_p)$ in tal caso è ortogonale (ossia la sua inversa è uguale alla sua trasposta): $V'V = VV' = I_p$. Nel caso quindi in cui la matrice A sia uguale alla matrice di covarianze oppure alla matrice di correlazione (essendo queste matrici simmetriche) la matrice V è sempre ortogonale.

1.3 Routines per il calcolo degli autovalori e degli autovettori

La risoluzione dell'equazione (1.14) in passato ha reso il problema del calcolo degli autovalori molto difficile. Evariste Galois¹⁰ (1811-1832) ha dimostrato che non esiste nessuna formula algebrica per calcolare le radici di un polinomio di grado $p \geq 5$, nel qual caso occorre utilizzare metodi numerici iterativi.

La funzione MATLAB per trovare i coefficienti del polinomio caratteristico si chiama `poly`. Ad esempio `polinomiocar=poly([2 2; 1 3])` restituisce il vettore riga `polinomiocar` di lunghezza 3 contenente i numeri (v. i coefficienti dell'equazione di secondo grado riportata nell'equazione 1.15):

$$1 - 5 \quad 4$$

La funzione `roots(polinomiocar)` consente di trovare le radici del polinomio caratteristico, ossia gli autovalori .

La function `eig` calcola direttamente tutti gli autovalori e gli autovettori della matrice A . Più precisamente, `e=eig(A)` fornisce nel vettore colonna di output e tutti gli autovalori della matrice A . Al contrario, `[V,D]=eig(A)` fornisce la matrice diagonale D , contenente gli autovalori sulla diagonale principale, e la matrice V , contenente gli autovettori.

Osservazione la funzione `eig` non sempre restituisce gli autovalori ordinati di conseguenza è necessario ordinare gli autovalori ed i corrispondenti

¹⁰<https://www.torinoscienza.it/personaggi/evariste-galois>

autovettori utilizzando la funzione `sort` in modo tale che l'elemento 1,1, di Λ contenga il più grande autovalore e la prima colonna di V l'autovettore corrispondente, l'elemento 2,2 contenga il secondo autovalore più grande e la seconda colonna di V l'autovettore corrispondente (v. ad esempio la sezione 1.4.1).

1.4 Scomposizione spettrale

Qualsiasi matrice quadrata simmetrica $S_{p \times p}$ può essere scomposta come segue:

$$S = V \Lambda V' \tag{1.17}$$

dove V è la matrice ortonormale ($V'V = I_p$) che contiene nelle colonne i p autovettori: $V = (v_1, v_2, \dots, v_p)$ e Λ è la matrice diagonale che contiene sulla diagonale principale i p autovalori di S . Nel linguaggio dell'algebra lineare si dice che le colonne della matrice V formano una base ortonormale.

S può anche essere scritta come:

$$\begin{aligned}
 S &= (v_1, v_2, \dots, v_p) \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix} \begin{pmatrix} v'_1, \\ v'_2 \\ \dots, \\ v'_p \end{pmatrix} \\
 &= (\lambda_1 v_1, \lambda_2 v_2, \dots, \lambda_p v_p) \begin{pmatrix} v'_1, \\ v'_2 \\ \dots, \\ v'_p \end{pmatrix} \\
 S &= \sum_{i=1}^p \lambda_i v_i v'_i
 \end{aligned}$$

Vediamo ora un esempio di scomposizione spettrale della matrice

$$A = \begin{pmatrix} \frac{1}{3} & \frac{\sqrt{20}}{3} \\ \frac{\sqrt{20}}{3} & \frac{2}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{2\sqrt{5}}{3} \\ \frac{2\sqrt{5}}{3} & \frac{2}{3} \end{pmatrix}.$$

Risolvendo l'equazione caratteristica: $\lambda_i^2 - \lambda_i - 2 = 0$, otteniamo che $\lambda_1 = 2$ e $\lambda_2 = -1$. I corrispondenti autovettori normalizzati sono riportati nella matrice che segue

$$V = \begin{pmatrix} \frac{2}{3} & -\frac{\sqrt{5}}{3} \\ \frac{\sqrt{5}}{3} & \frac{2}{3} \end{pmatrix}.$$

La scomposizione spettrale è la seguente:

$$A = V \Lambda V' = \begin{pmatrix} \frac{1}{3} & \frac{2\sqrt{5}}{3} \\ \frac{2\sqrt{5}}{3} & \frac{2}{3} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & -\frac{\sqrt{5}}{3} \\ \frac{\sqrt{5}}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & \frac{\sqrt{5}}{3} \\ -\frac{\sqrt{5}}{3} & \frac{2}{3} \end{pmatrix}.$$

1.4.1 La scomposizione spettrale attraverso il calcolo simbolico

L'istruzione

```
A=sym([1/3 sqrt(20)/3; sqrt(20)/3 2/3]);
```

produce

```
A =  
[      1/3, (2*5^(1/2))/3]  
[(2*5^(1/2))/3,      2/3]
```

La chiamata alla funzione eig

```
[V,Lambda]=eig(A);
```

produce

```
V =  
  
[-5^(1/2)/2, (2*5^(1/2))/5]  
[      1,      1]
```

```
Lambda =
```

```
[-1, 0]  
[ 0, 2]
```

Si noti che gli autovalori non ordinati in senso crescente e gli autovettori (nel calcolo simbolico) non sono normalizzati. In generale per effettuare il riordinamento in ordine non crescente si può procedere come segue:

```

d=diag(Lambda);
[~,permutation]=sort(d,"descend");

% Lambda1(1,1) autovalore più grande
% Lambda1(2,2) secondo autovalore più grande
Lambda1=diag(d(permutation));
% Prima colonna di V = primo autovettore associato
% all'autovalore più grande
% Seconda colonna di V = secondo autovettore associato
% al secondo autovalore più grande
V=V(:,permutation);

```

L'ultimo passaggio consiste fare in modo che la somma dei quadrati di ogni colonna della matrice degli autovettori sia pari a 1 (si divide quindi per la radice della somma dei quadrati degli elementi delle colonne)

```
V1=V./sqrt(sum(V.^2,1));
```

Le istruzioni

```

disp("Matrice degli autovalori ordinati (in senso non crescente)")
disp(Lambda1)
disp("Matrice degli autovettori (corrispondenti agli autov. ordinati)")
disp(V1)

```

producono

```

Matrice degli autovalori ordinati (in senso non crescente)
[2,  0]

```

$[0, -1]$

Matrice degli autovettori (corrispondenti agli autov. ordinati)

$\begin{bmatrix} 2/3, & -5^{(1/2)}/3 \end{bmatrix}$

$\begin{bmatrix} 5^{(1/2)}/3, & 2/3 \end{bmatrix}$

A questo punto si può controllare che la matrice $V1$ sia ortogonale e che $V1 * \text{Lambda1} * V1'$ ricostruisce la matrice A di partenza. Le istruzioni

```
disp('Verifico l''ortogonalità della matrice V1')
```

```
disp(V1'*V1)
```

producono

$[1, 0]$

$[0, 1]$

Le istruzioni

```
disp('Verifico la ricostruzione della matrice di partenza')
```

```
disp(V1*Lambda1*V1')
```

producono

$\begin{bmatrix} 2/3, & -(2*5^{(1/2)})/3 \end{bmatrix}$

$\begin{bmatrix} -(2*5^{(1/2)})/3, & 1/3 \end{bmatrix}$

1.5 Introduzione ai poligoni

L'obiettivo di questa sezione è:

1. mostrare come si aggiunge un ellisse di confidenza ad un diagramma di dispersione (v. Figura 1.6); Introdurre i concetti di assi principali dell'ellisse. L'asse maggiore dell'ellisse giace nella direzione di maggior variabilità. L'asse minore è ortogonale al precedente.
2. calcolare l'area ed il perimetro di un poligono¹¹;
3. trovare i punti che giacciono all'interno (all'esterno) di un determinato poligono (v. Figura 1.7);
4. effettuare l'intersezione tra due poligoni e trovare i punti dentro e fuori l'intersezione (v. Figura 1.8).

A titolo di esempio si considerino i dati sulla spesa pubblicitaria e sul fatturato di 25 aziende riportati nella Tabella 1.2. L'ellisse di confidenza al 95 per cento sovrapposto al diagramma di dispersione si costruisce chiamando la funzione `ellipse` di FSDA toolbox (v. Figura 1.6).

Esercizio. Caricare i dati presenti nel file `SpesaFatt.xlsx`. Creare il diagramma di dispersione dei punti. Aggiungere al diagramma di dispersione un ellisse di confidenza al 95 per cento mostrando gli assi principali dell'ellisse. Aggiungere la legenda.

Soluzione

```
% Caricamento in memoria dei dati contenuti nel file di Excel di input.  
Xtable=readtable("SpesaFatt.xlsx");
```

¹¹Un poligono è una figura geometrica corrispondente alla porzione di piano limitata da una linea spezzata chiusa non intrecciata.

Tabella 1.2: Spesa pubblicitaria e fatturato per 25 aziende del settore tessile (dati in milioni di Euro)

Spesa pubblicitaria	Fatturato
95	191
89	195
88	181
93	183
84	176
97	208
90	189
99	197
92	188
90	192
98	179
87	183
90	174
99	190
91	188
77	163
95	195
93	186
85	181
80	175
94	192
83	174
79	176
107	197
103	190

```
X=Xtable{:,:};  
  
% Calcolo del vettore riga contenente le medie aritmetiche delle  
% due variabili (il cosiddetto centroide)  
meaX=mean(X);  
  
% S = matrice di covarianze  
S=cov(X);  
  
  
% Diagramma di dispersione  
plot(X(:,1),X(:,2),'o','LineWidth',3)  
xlabel('X1=Spesa pubblicitaria (mln €)')  
ylabel('X2=Fatturato (mln €)')  
  
  
hold('on')  
  
% L'ellisse di confidenza viene aggiunto chiamando la funzione ellipse  
% specificando il vettore delle medie e la matrice di covarianze  
Ell= ellipse(meaX,S,0.95);  
axis equal  
  
  
% Aggiunta della legenda  
legend("Diagramma di dispersione",...  
      "Coordinate ellisse di confidenza al 95%",...  
      'Location','southeast')
```

L'output di questo codice è la Figura 1.6.

Esercizio: calcolare in maniera numerica l'area, il perimetro ed il centroide dell'ellisse mostrato nella Figura 1.6. Confrontare il centroide ottenuto in

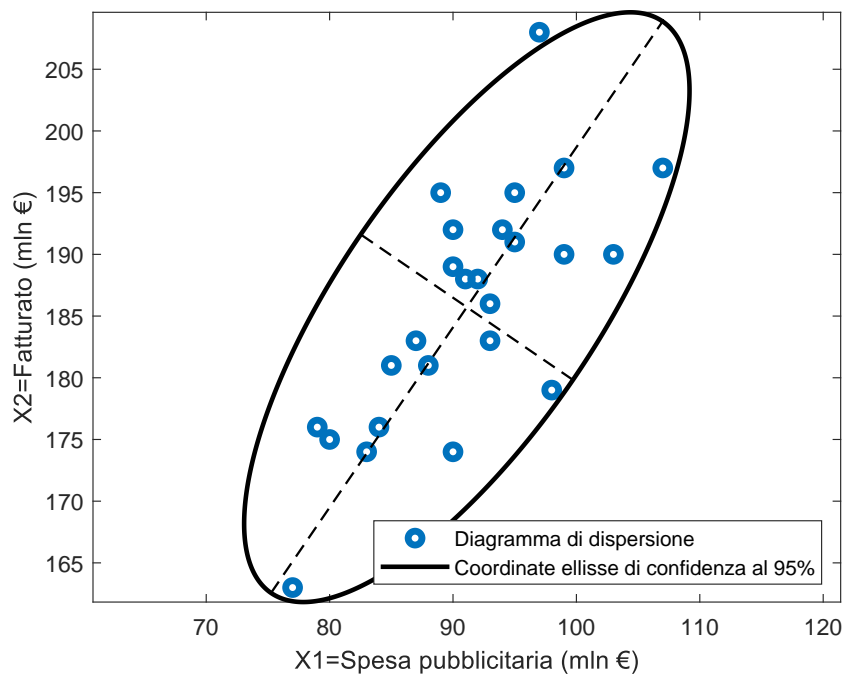


Figura 1.6: sovrapposizione di un ellisse di confidenza al diagramma di dispersione. L'asse maggiore dell'ellisse passa attraverso la direzione di massima variabilità dei punti proiettati.

maniera numerica con il vettore \bar{x} delle medie aritmetiche.

La funzione che consente di trasformazione un insieme di coordinate x, y in un poligono si chiama `polyshape`. All'output di questa funzione è sufficiente applicare le funzioni `perimeter`, `area` o `centroid` per calcolare rispettivamente il perimetro, l'area o il centroide del poligono¹².

Soluzione.

```
% Trasformo le coordinate dell'ellisse in un poligono
ellp=polyshape(Ell);
% disp("Area dell'ellisse (in maniera numerica)")
AreaEll=area(ellp);
```

¹²Si veda la sezione denominata *Geometric quantities* dell'help della funzione `polyshape`.


```
% disp("Perimetro dell'ellisse (in maniera numerica)")
PerimetroEll=perimeter(ellp);
disp("Centroide dell'ellisse (in maniera numerica)")
[meaX1,meaX2]=centroid(ellp);
disp([meaX1, meaX2])

disp("Centroide (medie aritmetiche) della matrice X")
disp(meaX)
```

L'output di questo codice produce

```
Centroide dell'ellisse (in maniera numerica)
    91.1200    185.7200

Centroide (medie aritmetiche) della matrice X
    91.1200    185.7200
```

c'è quindi una coincidenza tra il centroide calcolato in maniera numerica e quello vero. Nella sezione 3.6.3 vedremo come calcolare in maniera esatta la lunghezza dei semiassi dell'ellisse e la relativa area.

La funzione che consente di individuare i punti all'interno di un poligono si chiama **inpolygon**. La funzione che consente di aggiungere una retta con un determinata intercetta e coefficiente angolare si chiama **refline**. L'utilizzo di queste routines ha permesso di ottenere la Figura 1.7. La funzione che consente di effettuare l'intersezione tra due poligoni si chiama **intersect**.

Esercizio: trovare i punti del diagramma di dispersione dentro l'ellisse di confidenza. Aggiungere al grafico la retta di equazione $X_2 = 52.7355 +$

$1.4594X_1$. Vedremo successivamente (v. sezione 3.6.1) che questa è la retta che passa attraverso l'asse maggiore dell'ellisse.

Soluzione

```
% La funzion inpolygon restituisce un vettore booleano di lunghezza n
% contenente true per i punti che sono dentro l'ellisse
insideBoo = inpolygon(X(:,1),X(:,2),Ell(:,1),Ell(:,2));

figure
hold('on')
plot(X(insideBoo,1),X(insideBoo,2),'bo','LineWidth',3)
plot(X(~insideBoo,1),X(~insideBoo,2),'rx','LineWidth',3)
xlabel('X1=Spesa pubblicitaria (mln €)')
ylabel('X2=Fatturato (mln €)')
% Aggiunta dell'ellisse di confidenza
plot(Ell(:,1),Ell(:,2))
axis equal

% Aggiunta retta
% Viene aggiunta la retta con intercetta e pendenza richiesti
a=52.7355;
b= 1.4594;
refline(b,a)
legend("Punti dentro l'ellisse di confidenza", ...
      "Punti fuori dall'ellisse di confidenza",...
```

```
"Ellisse di confidenza","Retta che passa per l'asse maggiore dell'ellisse",...
'Location','southeast')
```

Il codice di cui sopra produce la Figura 1.7.

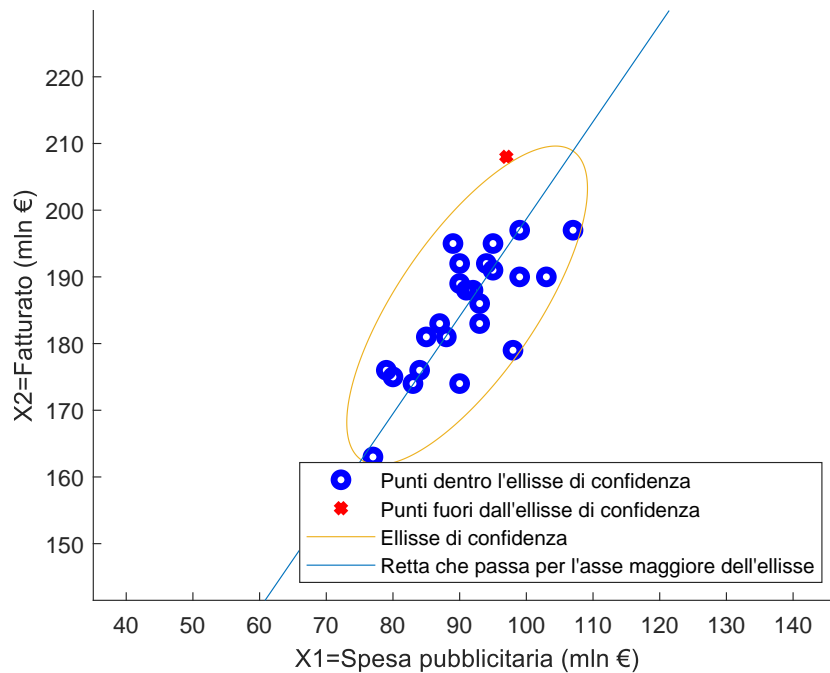


Figura 1.7: Punti dentro e fuori dall'ellisse di confidenza rappresentati con simboli e colori diversi. La retta che passa attraverso l'asse maggiore dell'ellisse è stata aggiunta.

Esercizio: trovare i punti della retta nel range $\min(x) - 20$ $\max(x) + 20$ che stanno dentro e fuori dall'ellisse di confidenza. Creare una nuova figura contenente i seguenti elementi: (v. per l'anteprima output la Figura 1.8) con

1. il diagramma di dispersione tra i punti.
2. l'ellisse di confidenza al 95 per cento;

3. i punti dentro e fuori dall'ellisse di confidenza rappresentati con simboli e colori diversi;
4. i punti della retta principale dentro e fuori dall'ellisse rappresentati con un diverso colore e un diverso stile di linea;
5. la legenda con l'indicazione delle diverse quantità.

Soluzione.

```
% Creazione delle coordinate x e y dei punti che stanno sulla retta
xcoo=[min(X(:,1))-20; max(X(:,1))+20];
ycoo=a+b*xcoo;
% La funzione intersect effettua l'intersezione tra due poligoni
% Il secondo argomento di input di intersect è una matrice a due
% colonne che definisce le coordinate della retta (bastano due righe)
[Pointsin,Pointsout]=intersect(ellp,[xcoo ycoo]);

% Creazione figura
% Diagramma di dispersione con punti dentro e fuori dall'ellisse di
% confidenza con simboli e colore diverso
figure
hold('on')
plot(X(insideBoo,1),X(insideBoo,2),'bo')
plot(X(~insideBoo,1),X(~insideBoo,2),'rx')
xlabel('X1=Spesa pubblicitaria (mln €)')
```

```

ylabel('X2=Fatturato (mln €)')

% Viene aggiunto al grafico l'ellisse di confidenza
plot(ellp)

% Punti della retta dentro e fuori l'ellisse di confidenza
plot(Pointsin(:,1),Pointsin(:,2),'b-.',Pointsout(:,1),Pointsout(:,2),'r')

% Aggiunta della legenda
legend("Punti dentro l'ellisse di confidenza", ...
      "Punti fuori dall'ellisse di confidenza",...
      "Ellisse di confidenza","Punti della retta dentro l'ellisse",...
      "Punti della retta fuori dall'ellisse",'Location','southeast')

axis equal

```

Questo codice consente di ottenere la Figura 1.8.

1.6 Proiezioni ortogonali

Dati due vettori x e y , il nostro obiettivo è trovare l'espressione del vettore \hat{x} che rappresenta la proiezione ortogonale¹³ di x in y (v. Figura 1.9). Ci chiediamo qual è l'espressione che definisce la lunghezza di \hat{x} ? Vedremo anche che il vettore \hat{x} minimizza la funzione $\|x - \hat{x}\|^2$.

Se θ è l'angolo tra x e y (v. Figura 1.9), la lunghezza della proiezione è data da

$$\|x\| |\cos \theta| = \|x\| \frac{|x'y|}{\|x\| \|y\|} = \frac{|x'y|}{\|y\|}. \quad (1.18)$$

¹³La parola ortogonale significa “ad angolo retto” ossia a 90 gradi.

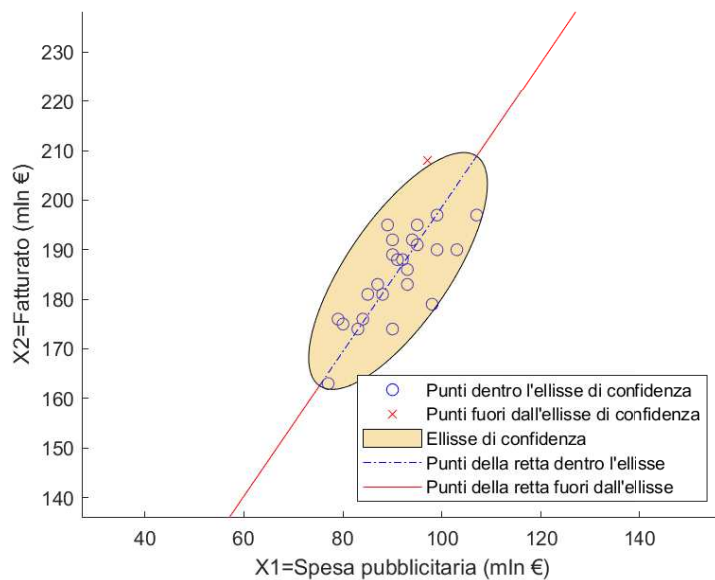


Figura 1.8: I punti della retta che passa per l'asse maggiore dell'ellisse che si trovano fuori dall'ellisse sono rappresentati con stile linea e colore diverso.

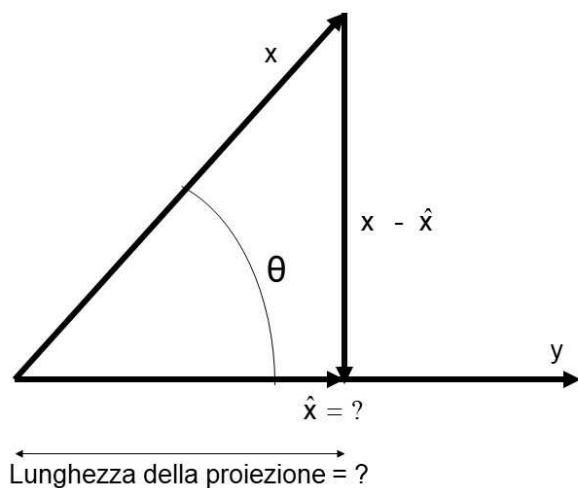
Il vettore \hat{x} che definisce la proiezione di x su y può essere scritto come $\hat{x} = ty$ dove t è un numero reale tale che $\|\hat{x}\| = \|ty\| = |t|\|y\| = \frac{|x'y|}{\|y\|}$ (si veda l'equazione 1.18). È facile controllare che

$$t = \frac{x'y}{\|y\|^2}. \quad (1.19)$$

Con questa scelta di t ,

$$\|\hat{x}\| = \|ty\| = |t|\|y\| = \frac{|x'y|}{\|y\|^2} \|y\| = \frac{|x'y|}{\|y\|}.$$

Riassumendo: il vettore che rappresenta la proiezione di x su y è funzione

Figura 1.9: Proiezione ortogonale del vettore x su y .

del prodotto scalare tra x e y ed è dato da

$$\hat{x} = \frac{x'y}{||y||} \frac{y}{||y||} = \frac{x'y}{y'y} y, \quad (1.20)$$

mentre la lunghezza della proiezione è data da

$$\frac{|x'y|}{||y||}. \quad (1.21)$$

Un altro modo per derivare l'espressione che definisce il vettore \hat{x} , si ottiene dopo aver notato che i vettori $x - \hat{x} = x - ty$ e $\hat{x} = ty$ sono ortogonali (si veda la Figura 1.10). Il requisito di ortogonalità implica che

$$(x - \hat{x})'\hat{x} = (x - ty)'ty = (x - ty)'y = 0. \quad (1.22)$$

Dall'equazione $(x - ty)'y = 0$ si trova che $t = \frac{x'y}{\|y\|^2}$.

Per mostrare che il vettore $\hat{x} = ty$ minimizza $\|x - \hat{x}\|^2$ dobbiamo calcolare il minimo rispetto a t , della funzione

$$f(t) = \|x - ty\|^2 = \|x\|^2 - 2tx'y + t^2\|y\|^2. \quad (1.23)$$

Differenziando rispetto a t si ottiene

$$f'(t) = -2x'y + 2t\|y\|^2.$$

Ponendo la precedente espressione a 0 si arriva alla stessa espressione di per t trovata in precedenza.

$$t = \frac{x'y}{\|y\|^2}.$$

Dato che $f''(t) = 2\|y\|^2 > 0$, il valore di t che è stato trovato corrisponde ad un minimo.

Esercizio.

Dato il punto x nello spazio \mathbb{R}^3 di coordinate $(1, 1, 0)$ e l'equazione esplicita della retta $r(\lambda) = r_0 + t v$ si determini il valore di t che minimizza la distanza del punto dalla retta ed il valore di questa distanza nel caso in cui $v = (1; -1; 0)'$ e $r_0 = (1.2; 2.6; 1.4)'$.

Soluzione.

Secondo la notazione della Figura 1.10, il punto di coordinate $(1, 1, 0)$ è il vettore x . L'equazione della retta $r(t) = r_0 + t v$ corrisponde al vettore y

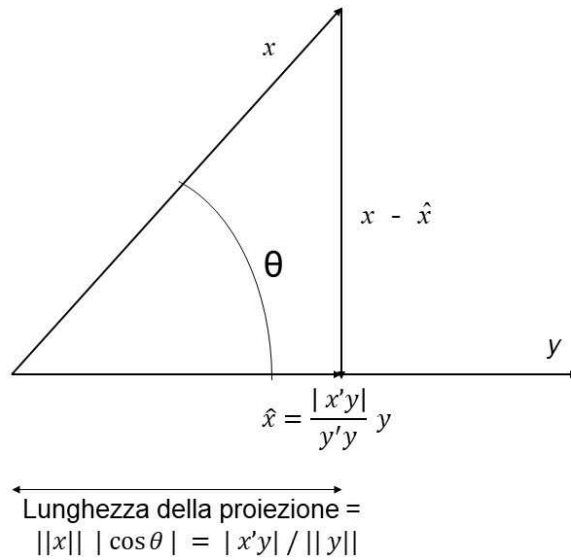


Figura 1.10: Proiezione ortogonale del vettore x su y con indicazione della lunghezza della proiezione.

nella Figura 1.10. Tenendo presente l'equazione (1.19) otteniamo

$$t = \frac{|(x - r_0)'v|}{v'v}$$

Dato che nel nostro caso la retta passa per l'origine $r_0 = (0; 0 \ 0)'$.

In istruzione MATLAB

```
% x = coordinate del vettore (punto) da proiettare su v
x=[1; 1; 0];
% La direzione del vettore v è
v=[1; -1; 0];
%t = x'*v/(v'*v)
```

```

t =x'*v/(v'*v);
r_0=[0; 0; 0];
hatx= r_0+t*v;
% dist= || x- hatx||
dist=norm(x-hatx);
disp(['La distanza è uguale a: ' num2str(dist)])

```

Il codice di cui sopra produce il seguente output:

```
La distanza è uguale a: 1.4142
```

1.7 L'espansione implicita

Abbiamo visto che le regole di algebra lineare implicano che due matrici devono avere lo stesso numero di righe e colonne per essere sommate, e che le matrici devono essere conformabili per effettuare la moltiplicazione. A partire dalla versione 2016b di MATLAB è stata introdotta la cosiddetta espansione implicita. In altri termini è possibile a partire dalla versione 2016b effettuare operazioni tra gli array purché abbiano “dimensioni compatibili”. Le dimensioni si dicono compatibili quando una dimensione in un array è uguale alla dimensione nell'altro array oppure è uguale ad 1. In altre parole, le dimensioni che compaiono in una matrice e non compaiono nell'altra, sono implicitamente idonee per l'espansione automatica. Seconda questa regola, ad esempio, se X è una matrice $n \times p$ e \bar{x}' è il vettore delle $1 \times p$ delle medie aritmetiche di X ottenuto tramite l'istruzione `overlinex =mean(X,1)`, $\bar{x}' =$

$(\overline{x}_1, \overline{x}_2, \dots, \overline{x}_p)$, la semplice istruzione

`Xtilde=X-overlinex`

espande automaticamente il vettore riga di lunghezza p \overline{x} in una matrice $n \times p$ come segue

$$\begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \dots & \overline{x}_p \\ \overline{x}_1 & \overline{x}_2 & \dots & \overline{x}_p \\ \dots & \dots & \dots & \dots \\ \overline{x}_1 & \overline{x}_2 & \dots & \overline{x}_p \end{pmatrix}$$

e crea la matrice degli scostamenti dalla media \tilde{X} come segue

$$\begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \dots & x_{1p} - \overline{x}_p \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \dots & x_{2p} - \overline{x}_p \\ \dots & \dots & \dots & \dots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \dots & x_{np} - \overline{x}_p \end{pmatrix}$$

Similmente l'istruzione

`Z=Xtilde./sigmaX`

dove il vettore riga `sigmaX` contiene gli scostamenti quadratici medi delle variabili originarie ottenuto come `sigmaX=std(X)`, espande `sigmaX` in una

matrice $n \times p$ come segue

$$\begin{pmatrix} \sigma_1 & \sigma_2 & \dots & \sigma_p \\ \sigma_1 & \sigma_2 & \dots & \sigma_p \\ \dots & \dots & \dots & \dots \\ \sigma_1 & \sigma_2 & \dots & \sigma_p \end{pmatrix}$$

e crea la matrice degli scostamenti standardizzati come segue

$$Z = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{12}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{1p}-\bar{x}_p}{\sigma_p} \\ \frac{x_{21}-\bar{x}_1}{\sigma_1} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{2p}-\bar{x}_p}{\sigma_p} \\ \dots & \dots & \dots & \dots \\ \frac{x_{n1}-\bar{x}_1}{\sigma_1} & \frac{x_{n2}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{np}-\bar{x}_p}{\sigma_p} \end{pmatrix}$$

Secondo questa logica, se x è un vettore colonna $n \times 1$ e y è un vettore colonna di dimensione p , l'istruzione $x + y'$ produce una matrice di dimensione $n \times p$ definita come segue

$$x + y' = \begin{pmatrix} x_1 + y_1 & x_1 + y_2 & \dots & x_1 + y_p \\ x_2 + y_1 & x_2 + y_2 & \dots & x_2 + y_p \\ \dots & \dots & \dots & \dots \\ x_n + y_1 & x_n + y_2 & \dots & x_n + y_p \end{pmatrix}$$

Esercizio.

Per la replicabilità dei risultati utilizzare il seed 123. Generare un matrice di numeri casuali di dimensione 20×3 denominata X dalla distribuzione $N(10, 3)$; Contaminare le osservazioni 5 e 8 con il valore 100.

1. creare la matrice degli scostamenti standardizzati Z , utilizzando l'espansione implicita e verificare il risultato con quello della funzione `zscore`. Rappresentare tramite barre la matrice degli scostamenti standardizzati e commentare i risultati.
2. creare la matrice degli scostamenti standardizzati robusti utilizzando la mediana ed il MAD normalizzato ($\text{MAD} \times 1.4826$) e verificare il risultato con la funzione `zscoreFS` di FSDA toolbox. Rappresentare tramite barre la matrice degli scostamenti standardizzati e commentare i risultati.

Soluzione:

```
n=20;
p=3;
rng(123)
X=10+3*randn(n,p);
% Un modo alternativo per generare i dati era
%X=normrnd(3,10,n,p);
% Contaminazione delle righe 5 e 8 con il valore 100
X([5 8],:)=100;
% Z = matrice degli scostamenti standardizzati utilizzando
% l'espansione implicita
Z=(X-mean(X))./std(X);
% Zchk = matrice degli scostamenti standardizzati utilizzando
% la funzione zscore
```

```

Zchk=zscore(X);
maxdiff=max(abs(Z-Zchk),[],"all");
assert(maxdiff<1e-12,"Errore di programmazione nella costruzione " + ...
      "della matrice degli scostamenti standardizzati")

bar(Z)

title(['Rappresentazione tramite barre ' ...
      'degli scostamenti standardizzati'])

% Zrob = matrice degli scostamenti standardizzati
% robusta utilizzando l'espansione implicita
Zrob=(X-median(X)) ./ (1.4826* mad(X,1));
% Zrobchk = matrice degli scostamenti standardizzati
% robusta utilizzando la funzione zscoreFS
Zrobchk=zscoreFS(X);
maxdiff=max(abs(Zrob-Zrobchk),[],"all");
assert(maxdiff<1e-4,"Errore di programmazione nella costruzione " + ...
      "della matrice degli scostamenti standardizzati robusti")

figure

bar(Zrob)

title(['Rappresentazione tramite barre ' ...
      'degli scostamenti standardizzati robusti'])

```

I due grafici prodotti sono riportati nelle Figure 1.11 e 1.12.

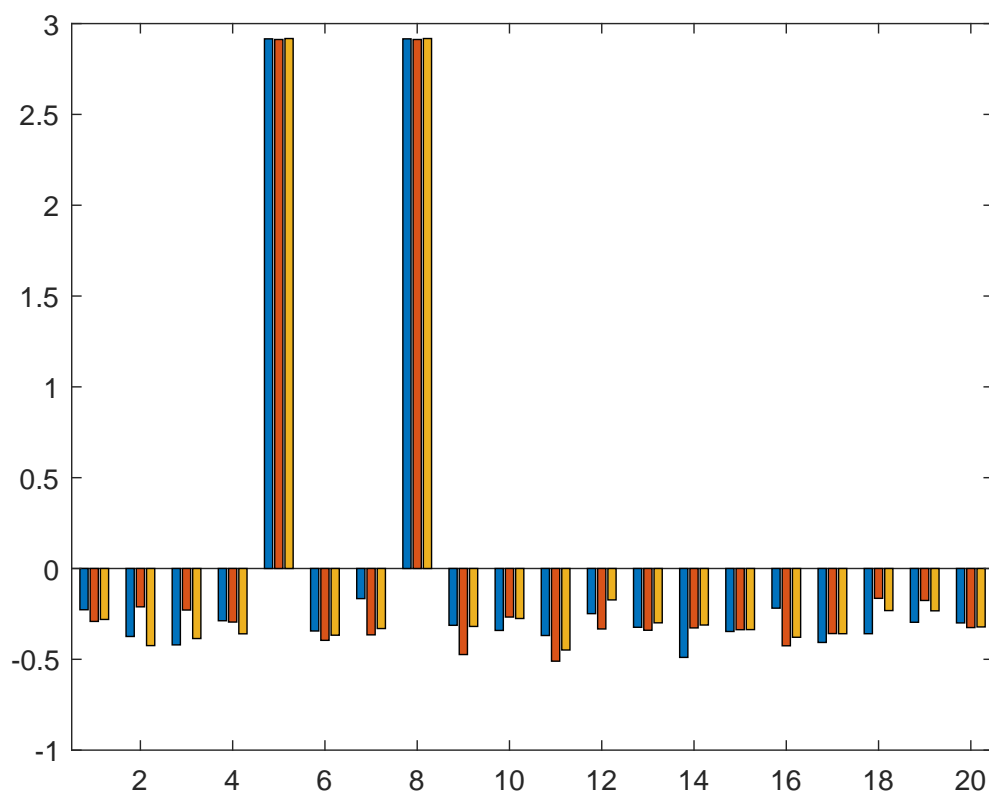


Figura 1.11: Rappresentazione tramite barre degli scostamenti standardizzati. A causa della presenza dei due valori anomali gli scostamenti standardizzati di tutte le altre unità sono negativi per rispettare il vincolo di somma a zero.

1.8 Matrice di varianze e correlazione tramite espressioni matriciali

L'obiettivo di questa sezione è mostrare che, tutte le matrici che abbiamo visto finora: matrice degli scostamenti standardizzati, matrice di covarianze, matrice di correlazione, possono essere ottenute tramite semplici calcoli matriciali. Queste nozioni sono importanti per poter capire i dettagli matematici delle tecniche multivariate di analisi in componenti principali, analisi delle corrispondenze e cluster analysis che vedremo nei capitoli successivi.

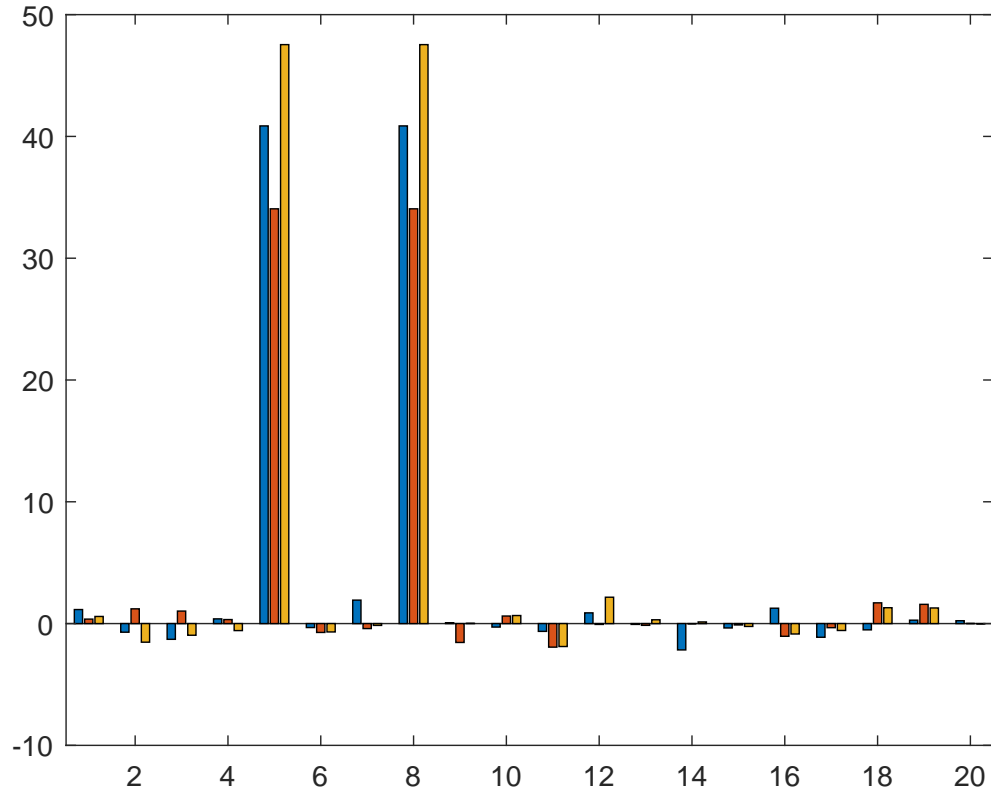


Figura 1.12: Rappresentazione tramite barre degli scostamenti standardizzati robusti. La standardizzazione robusta non risente della presenza di valori anomali.

Data $X_{n \times p}$ = matrice dei dati originale, il vettore $1 \times p$ che contiene la somma degli elementi di ogni colonna:

$$\left(\sum_{i=1}^n x_{i1} \quad \sum_{i=1}^n x_{i2}, \dots, \sum_{i=1}^n x_{ip} \right)$$

può essere scritto come $(1_{n \times 1})'X$ dove $1_{n \times 1}$ = vettore di lunghezza n con tutti gli elementi uguali ad 1.

La matrice degli scostamenti dalla media \tilde{X} , in termini matriciali può essere ottenuta come:

$$\tilde{X} = HX$$

dove

$$H = I_n - \frac{1}{n} 1_{n \times 1} (1_{n \times 1})' = I_n - \frac{1_{n \times n}}{n}$$

(definita spesso come centering matrix) è una matrice simmetrica e idempotente che presenta rango e traccia uguale a $n - 1$.

Per esempio:

$$H_1 = 0; \quad H_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$H_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

Quando facciamo il prodotto di H_n per un vettore $v = (v_1, \dots, v_n)'$ otteniamo:

$$H_n v = v - \frac{1}{n} (1_{n \times 1} (1_{n \times 1})') v = v - \frac{1}{n} \sum_{i=1}^n v_i 1_{n \times 1} = v - \bar{v} 1_{n \times 1}$$

dove \bar{v} è la media della componenti del vettore v :

$$\bar{v} = \frac{1}{n} (1_{n \times 1})' v = \frac{1}{n} \sum_{i=1}^n v_i.$$

Si vede dunque chiaramente che la media del vettore v è stata sottratta al vettore risultante $H_n v$. Si noti inoltre che: $H = H'$ e $H = H \times H$, $tr(H) = n - tr(1_{n \times 1} (1_{n \times 1})')/n = n - tr(1_{1 \times n} 1_{n \times 1})/n = n - tr(n/n) = n - 1$.

Il vettore riga che contiene la somma degli scostamenti dalla media

$$\left(\sum_{i=1}^n (x_{i1} - \bar{x}_1), \sum_{i=1}^n (x_{i2} - \bar{x}_2), \dots, \sum_{i=1}^n (x_{ip} - \bar{x}_p) \right)$$

può essere scritto come:

$$(1_{n \times 1})' \tilde{X} = 0_{1 \times p}$$

dove $0_{1 \times p}$ è un vettore riga con elementi tutti uguali a zero di lunghezza p .

La matrice di covarianze S può essere scritta come

$$\begin{aligned} S_{p \times p} &= \tilde{X}' \tilde{X} / (n-1) = X' H X / (n-1) \\ &= (\tilde{x}_1 \ \tilde{x}_2 \ \dots \ \tilde{x}_n) \begin{pmatrix} \tilde{x}'_1 \\ \tilde{x}'_2 \\ \dots \\ \tilde{x}'_n \end{pmatrix} / (n-1) = \sum_{i=1}^n \tilde{x}_i \tilde{x}'_i / (n-1) \end{aligned}$$

Data $D_{p \times p}$ = matrice diagonale che contiene sulla diagonale principale gli scostamenti quadratici medi (campionari) delle variabili originarie $D = \text{diag}(s_1, \dots, s_p)$, la matrice $Z_{n \times p}$ che contiene gli scostamenti standardizzati può essere scritta come:

$$Z = \tilde{X} D^{-1}$$

La matrice di correlazione si può scrivere come

$$R = Z' Z / (n-1) = D^{-1} \tilde{X}' \tilde{X} D^{-1} / (n-1) = D^{-1} X' H X D^{-1} / (n-1)$$

Data la matrice di covarianze S , la matrice di correlazione, in termini matri-

ciali può essere scritta come:

$$R = D^{-1}SD^{-1}$$

Esercizio Data una matrice di dati di dimensione $n \times p$ costruire

1. la matrice H .
2. la matrice degli scostamenti dalla media e degli scostamenti standardizzati \tilde{X} utilizzando H
3. la matrice di covarianze S utilizzando \tilde{X}
4. la matrice $D_{p \times p}$, matrice diagonale che contiene sulla diagonale principale gli scostamenti quadratici medi (campionari) delle variabili originarie
5. la matrice di correlazione tramite le formule $R = Z'Z/(n-1) = D^{-1}X'HXD^{-1}/(n-1) = D^{-1}SD^{-1}$

Soluzione.

$n=10;$

$p=5;$

$X=\text{randn}(n,p);$

```

uno=ones(n,1);

% H = centering matrix. Matrice idempotente e simmetrica che consente di
% passare dalla matrice originaria alla matrice degli scostamenti
% standardizzati
H=eye(n)-uno*uno'/n;
Xtilde=H*X;

%% 9) Matrice di covarianze in maniera matriciale
S=Xtilde'*Xtilde/(n-1);
disp('Matrice di covarianze tramite la matrice Xtilde')
disp(S)

Schk=X'*H*X/(n-1);
disp('Matrice di covarianze ottenuta come X''H X/(n-1)')
disp(Schk)
disp('Matrice di covarianze ottenuta direttamente tramite la funzione cov')
disp(cov(X))

%% Matrice degli scostamenti standardizzati in maniera matriciale
sigmas=sqrt(diag(S));
D=diag(sigmas);
% invD=inv(D);
invD=D^-1;
Z=H*X*invD;

```

```

disp('Matrice degli scostamenti standardizzati in maniera matriciale')
disp(Z)
disp('Matrice degli scostamenti standardizzati tramite la funzione zscore')
Zchk=zscore(X);
disp(Zchk)

%% Matrice di correlazione in maniera matriciale.
sigmas=sqrt(diag(S));
R=Z'*Z/(n-1);
disp("Matrice di correlazione ottenuta come Z'Z/(n-1)")
disp(R)
disp('Matrice di correlazione tramite la funzione corr')
Rchk=corr(X);
disp(Rchk)
disp('Matrice di correlazione tramite le matrici D, X H ')
Rchk1=invD*X' * H * X *invD /(n-1);
disp(Rchk1)

```


Capitolo 2

Le distanze e gli indici di similarità

In questo capitolo analizziamo le “prossimità” tra unità statistiche, alle quali corrispondono i vettori riga x'_i , ($i = 1, \dots, n$) nella matrice dei dati. Un indice di prossimità tra due generiche righe unità i e j è definito come funzione dei rispettivi vettori riga della matrice dei dati:

$$f(x'_i, x'_j)$$

Vedremo di seguito una serie di esempi dove si specifica la natura di f .

Osservazione: con il termine prossimità ci si riferisce sia al concetto di rassomiglianza tra e unità sia a quello antitetico di diversità dato che è equivalente affermare che due unità sono molto simili oppure poco diverse.

Le informazioni fornite tra gli indici di prossimità tra coppie di elementi costituiscono la premessa per l'individuazione di gruppi di unità omogenee.

La formazione di gruppi omogenei di unità (che sarà oggetto del capitolo sulla *cluster analysis*) può interpretarsi come una riduzione delle dimensioni dallo spazio \mathbb{R}^n , poiché si riuniscono le unità in k sottoinsiemi (tipicamente con $k \ll n$).

2.1 Definizione di distanze

Si dice distanza tra due punti corrispondenti ai vettori x e $y \in \mathbb{R}^p$, una funzione che gode delle seguenti proprietà:

1. non negatività

$$d(x, y) \geq 0 \quad \forall x, y \in \mathbb{R}^p$$

2. identità

$$d(x, y) = 0 \quad \text{se e solo se } x = y$$

3. simmetria

$$d(x, y) = d(y, x) \quad \forall x, y \in \mathbb{R}^p$$

4. disuguaglianza triangolare

$$d(x, y) \leq d(x, z) + d(y, z) \quad \forall x, y \in \mathbb{R}^p \quad (2.1)$$

Uno spazio con riferimento al quale è definita una distanza è detto spazio metrico.

In presenza di una matrice dei dati con variabili tutte quantitative, la distanza (che corrisponde ad una categoria particolare degli indici di prossimità introdotti nella precedente sezione) è calcolata sui vettori riga x'_i e x'_j ed è indicata come segue:

$$d(x'_i, x'_j) = d_{ij}$$

2.2 Alcuni tipi di distanza

Il caso più noto di distanza tra due punti è la distanza Euclidea.

Definizione: si dice distanza Euclidea tra due unità statistiche i e j la norma euclidea (v. sezione 1.1.5) della differenza tra i rispettivi vettori.

$${}_2d_{ij} = \|x_i - x_j\| = \sqrt{(x_i - x_j)'(x_i - x_j)} = \sqrt{\sum_{s=1}^p (x_{is} - x_{js})^2}$$

Nel caso di due sole variabili X_1 e X_2 è possibile rappresentare nel piano cartesiano i punti corrispondenti alle unità statistiche. La distanza Euclidea è uguale alla lunghezza del segmento che li unisce (v. Figura 2.1). Più precisamente, siano $x'_i = (x_{i1}, x_{i2})$ e $x'_j = (x_{j1}, x_{j2})$ i vettori corrispondenti a due generiche unità statistiche. La distanza Euclidea tra essi è

$${}_2d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

Questa espressione è la radice quadrata della somma dei quadrati costruiti sui cateti del triangolo rettangolo che compare nella Figura 2.1, che per il

teorema di Pitagora, equivale all'ipotenusa del triangolo stesso. La denominazione *spazio Euclideo* è un modo formale di chiamare lo spazio fisico a cui siamo abituati e che è stato tacitamente assunto nella rappresentazione dei punti dei grafici precedenti. Gli spazi Euclidei a bassa dimensione sono gli spazi geometrici con cui siamo soliti convivere: ad una dimensione abbiamo una linea, a due dimensioni abbiamo un piano e a 3 dimensioni abbiamo lo spazio 3D che vediamo intorno a noi.

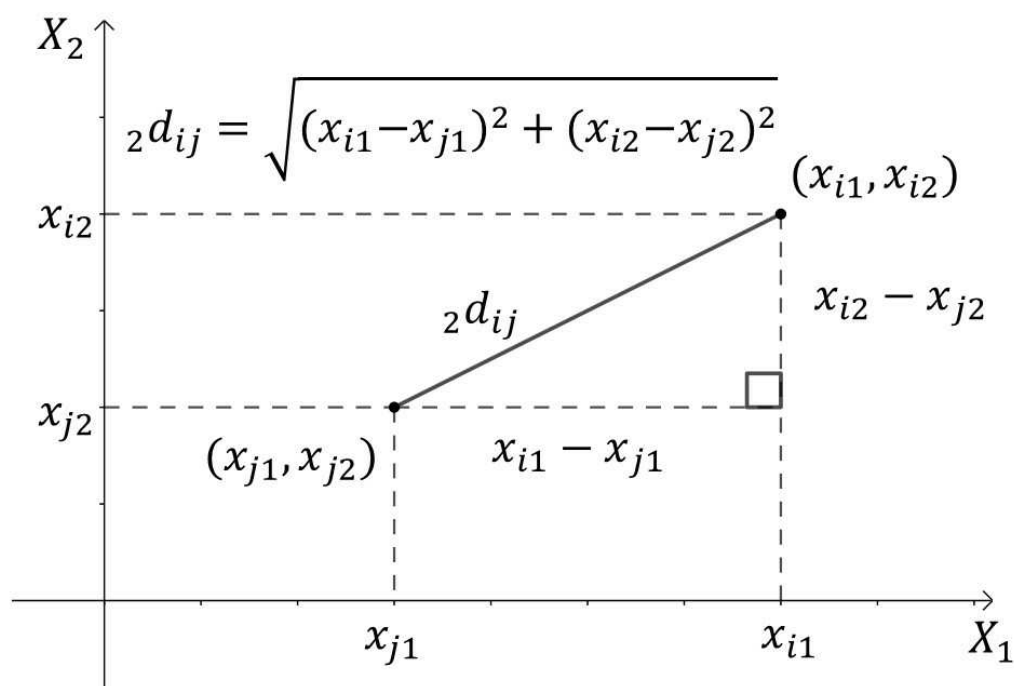


Figura 2.1: Distanza Euclidea tra due punti di coordinate $x'_i = (x_{i1}, x_{i2})$ e $x'_j = (x_{j1}, x_{j2})$.

Un altro tipo di distanza di notevole interesse è fornito dalla seguente:

Definizione: si dice distanza della città a blocchi (*city-block distance*) tra

due unità l'espressione:

$${}_1d_{ij} = \sum_{s=1}^p |x_{is} - x_{js}|$$

Nella Figura 2.1 questa distanza corrisponde alla somma dei due cateti ($|x_{i2} - x_{j2}|$ e $|x_{i1} - x_{j1}|$) ed il nome le deriva dal fatto che essa è la lunghezza che si deve percorrere per spostarsi da x_i a x_j , qualora sia consentito muoversi solo nelle direzioni parallele agli assi, come avviene in una città con una griglia regolare di strade che si intersecano ad angolo retto (*city block*).

I due tipo di distanza precedenti possono ottenersi come casi particolari da una formula più generale.

Definizione: si dice distanza di Minkowski di ordine k tra le unità i e j la norma k -esima (eq. 1.2) della differenza tra i due vettori:

$${}_kd_{ij} = \|x_i - x_j\|_k = \left(\sum_{s=1}^p |x_{is} - x_{js}|^k \right)^{1/k} \quad k > 0$$

Si ricava facilmente che la metrica di Minkowski per $k = 1$ coincide con la distanza *cityblock* e per $k = 2$ equivale alla distanza Euclidea. Tutto questo giustifica la simbologia ${}_1d_{ij}$ e ${}_2d_{ij}$ adottata in precedenza.

Inoltre:

$$\lim_{k \rightarrow \infty} {}_kd_{ij} = \max_j |x_{is} - x_{js}|$$

definisce la distanza chiamata lagrangiana o distanza di Chebychev o distanza della scacchiera (v. Figura 2.2). Il nome della distanza della scacchiera deriva dalle mosse necessarie al re per spostarsi da una casella ad un'altra della scacchiera¹.

¹Nel gioco degli scacchi il re si può muovere in una delle case adiacenti (anche diagonalmente) a quella occupata.


	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

Figura 2.2: Distanza di Chebychev o distanza della scacchiera. I numeri nella figura indicano il numero di mosse necessarie al re per arrivare in quella determinata casella. Questo numero di mosse è sempre pari al massimo tra le differenze delle due coordinate orizzontali e verticali.

Si noti che nel caso di una sola variabile X_1 la metrica di Minkowski per qualunque valore di k risulta uguale a:

$$d_{ij} = |x_{i1} - x_{j1}|$$

ossia alla distanza tra due punti in \mathbb{R}^1 .

Osservazione: la funzione `pdist` chiamata con un solo argomento di input consente di calcolare le distanze Euclidee tra le righe della matrice dei dati. Il secondo argomento di input di `pdist` consente di specificare il tipo di distanza che si vuole calcolare.

Esercizio: la matrice dei dati riportata nella Tabella 2.1 riporta il numero

di certificazioni ISO 9000 in cinque Paesi in un orizzonte temporale di due anni consecutivi.

Tabella 2.1: Matrice dei dati di partenza X . Numero di certificazioni ISO 9000 in 5 paesi in un periodo di 2 anni (dati in migliaia).

	Anno 1	Anno 2
A	1.2	2.1
B	3.4	5.1
C	1.9	2.2
D	4.1	2.3
E	4.2	2.9

I comandi

```
X = [1.2 2.1; 3.4 5.1; 1.9 2.2; 4.1 2.3; 4.2 2.9];
```

```
pdist(X)
```

```
producono
```

```
Columns 1 through 6
```

```
3.7202    0.7071    2.9069    3.1048    3.2650    2.8862
```

```
Columns 7 through 10
```

```
2.3409    2.2023    2.4042    0.6083
```

Il primo numero è la distanza Euclidea tra il paese A ed il paese B , il secondo numero è la distanza Euclidea tra il paese A ed il paese C . L'ultimo numero è la distanza Euclidea tra il paese D ed il paese E . È possibile formattare l'output in una matrice in cui sia le righe sia le colonne corrispondono alle righe della matrice dei dati (in questo caso i 5 paesi) e un suo elemento generico corrisponde alla distanza misurata tra il Paese sulla riga i e il Paese sulla colonna j . Questa matrice è detta matrice delle distanze. A tale scopo si utilizza il comando `squareform` digitando

```
D=squareform(pdist(X))
```

Questo comando produce

```
D =
```

0	3.7202	0.7071	2.9069	3.1048
3.7202	0	3.2650	2.8862	2.3409
0.7071	3.2650	0	2.2023	2.4042
2.9069	2.8862	2.2023	0	0.6083
3.1048	2.3409	2.4042	0.6083	0

Il codice completo per ottenere la matrice delle distanze in formato table è riportato di seguito.

```
% Calcolo della matrice delle distanze euclidee
D=squareform(pdist(X));
% nomirighe = vettore colonne che contiene le
% etichette dei paesi in formato string.
nomirighe=string(('A':'E')));
% Avvertenza se non chiamiamo la funzione string
% quando si va a chiamare array2table otteniamo l'errore che segue
% The RowNames property must be a string array or a cell array,
% with each name containing one or more characters.
% L'array D viene trasformato in table
Dtable=array2table(D,"RowNames",nomirighe,"VariableNames",nomirighe);
disp(Dtable)
```

Questo codice produce

	A	B	C	D	E
	-----	-----	-----	-----	-----
A	0	3.7202	0.70711	2.9069	3.1048
B	3.7202	0	3.265	2.8862	2.3409
C	0.70711	3.265	0	2.2023	2.4042
D	2.9069	2.8862	2.2023	0	0.60828
E	3.1048	2.3409	2.4042	0.60828	0

La matrice (table) delle distanze *cityblock* può essere ottenuta chiamando la funzione `pdist` con due argomenti di input. Se si rimpiazza nel codice sopra l'istruzione `D=squareform(pdist(X));` con

`D=squareform(pdist(X,"cityblock"));`

si ottiene la table riferita alle distanze *cityblock*.

	A	B	C	D	E
	---	---	---	---	---
A	0	5.2	0.8	3.1	3.8
B	5.2	0	4.4	3.5	3
C	0.8	4.4	0	2.3	3
D	3.1	3.5	2.3	0	0.7
E	3.8	3	3	0.7	0

Se il secondo argomento di input è `"minkowski"`, allora è possibile nel quarto argomento di input specificare il valore di k . Ad esempio, l'istruzione `D=squareform(pdist(X,"minkowski",3))` nel codice di cui sopra consente di ottenere.

	A	B	C	D	E
	-----	-----	-----	-----	-----
A	0	3.3516	0.70068	2.9003	3.0188
B	3.3516	0	3.028	2.8145	2.2347
C	0.70068	3.028	0	2.2001	2.3214
D	2.9003	2.8145	2.2001	0	0.60092
E	3.0188	2.2347	2.3214	0.60092	0

Dalle precedenti matrici si evince che il paese D è molto simile al paese E e anche che il paese A è molto simile al paese C . Al contrario, la coppia di paesi più distante in termini di certificazioni ISO è data da (A,B).

Esercizio: trovare il luogo dei punti distanti r dall'origine degli assi utilizzando i seguenti valori di k della distanza di Minkowski 0.25, 0.5, 1 (city-block), 2 (Euclidea) 4 e 10. Mostrare graficamente i contorni utilizzando una finestra grafica con 6 pannelli. In ogni pannello inserire nel titolo il valore di k che è stato utilizzato.

Soluzione.

Data l'equazione $|x|^k + |y|^k = r^k$, si tratta di trovare le coordinate y in corrispondenza delle coordinate x nell'intervallo $[-r, r]$

```
% Valore a piacere per r
```

```
r=2;
```

```
% Sequenza di valori di k richiesti dal testo dell'esercizio
```

```
kk=[0.25 0.5 1 2 4 10];
```

```
% Le coordinate x vanno da -r ad r
```



```

x=(-r:0.001:r)';

for j=1:length(kk)
    subplot(2,3,j)
    k=kk(j);
    % Trovo la coordinata y dell'equazione |x|^k+|y|^k=r^k
    y=(r^k-abs(x).^k).^(1/k);
    plot([x;x],[y;-y])
    % Vengono aggiunti gli assi cartesiani
    xline(0)
    yline(0)
    % Viene fissato il limite min e max per l'asse x
    xlim([-r*1.1 r*1.1])
    % Il valore di k utilizzato viene inserito nel titolo
    title(['k=' num2str(k)])
    % Stessa scala per i due assi
    axis equal
end

```

Il codice di cui sopra produce l'output mostrato nella Figura 2.3.

La metrica di Minkowski gode di determinate proprietà

Proprietà 1. La metrica di Minkowski è funzione decrescente dell'indice k per cui valgono le seguenti disuguaglianze:

$${}_1d_{ij} \geq {}_2d_{ij} \geq \cdots \geq {}_\infty d_{ij}$$

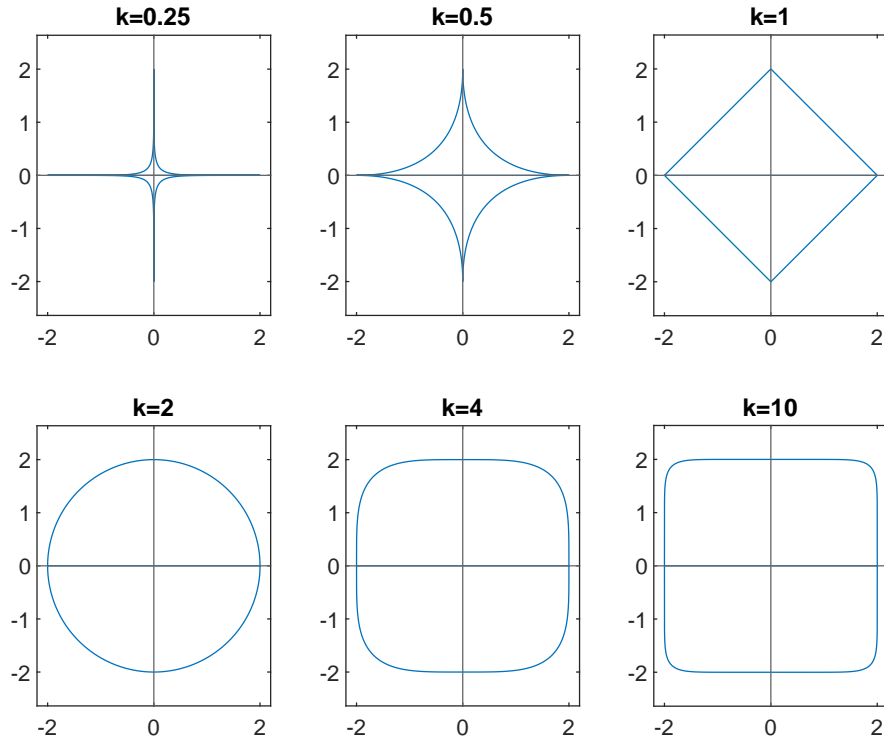


Figura 2.3: Luogo dei punti distanti $r = 2$ secondo la metrica di Minkowski per diversi valori di k . Secondo la distanza euclidea il contorno è il cerchio di raggio r . Secondo la distanza *cityblock* è il quadrato inscritto nella circonferenza. Secondo la distanza Lagrangiana è il quadrato circoscritto alla circonferenza. All'aumentare del valore di k il contorno dei punti equidistanti si avvicina sempre di più al quadrato circoscritto.

L'uguaglianza si verifica solo nel caso banale di $p = 1$ (dati unidimensionali).

Osservazione: le relazioni precedenti si ricavano dalla disuguaglianza di Jensen

$$\left(\sum_{s=1}^p x_s^k \right)^{1/k} \geq \left(\sum_{s=1}^p x_s^h \right)^{1/h} \quad h > k > 0 \quad x_s \geq 0$$

Proprietà 2. La metrica di Minkowski è invariante per traslazione delle variabili.

$${}_k d(x_i + c, x_j + c) = {}_k d(x_i, x_j)$$

dove c in questo caso indica un vettore p dimensionale di costanti note.

Proprietà 3: la distanza euclidea è invariante per trasformazioni ortogonali (rotazioni) delle variabili.

Questa proprietà si dimostra facilmente facendo riferimento ai vettori elementari e_i e_j di lunghezza n introdotti nella sezione 1.1.6. Si ha che

$$e_i'X = x_i' \quad x_i = X'e_i \quad x_i - x_j = X'e_i - X'e_j = X'(e_i - e_j)$$

quindi la distanza Euclidea al quadrato in funzione dei vettori elementari può essere scritta come

$${}_2d_{ij}^2 = (x_i - x_j)'(x_i - x_j) = (e_i - e_j)'XX'(e_i - e_j) \quad (2.2)$$

Utilizzando questa scrittura è immediato verificare che la moltiplicazione per una matrice ortogonale T (v. sezione 1.1.9) lascia invariato il valore della distanza Euclidea.

$$\begin{aligned} {}_2d_{ij}^2 &= (e_i - e_j)'(XT)(XT)'(e_i - e_j) \\ &= (e_i - e_j)'(XTT'X')(e_i - e_j) \\ &= (e_i - e_j)'XX'(e_i - e_j) \end{aligned}$$

Queste proprietà implicano che le distanze euclidee tra i punti in \mathbb{R}^p non mutano quando si effettua una traslazione degli assi di riferimento (ad esempio si passa da X a \tilde{X}) oppure quando si opera una rotazione degli stessi (si passa da \tilde{X} a $\tilde{X}T$ con T matrice ortogonale).

Osservazione: nel seguito di questo testo per denotare la distanza Euclidea utilizziamo il simbolo sia il simbolo completo ${}_2d_{ij}$ sia il simbolo più semplice d_{ij} . In altri termini, se il valore di k nella distanza di Minkowski è omesso si deve sempre intendere $k = 2$.

2.3 Gli indici di distanza e gli indici di dissimilarità

Si dice indice di distanza tra due vettori $x, y \in \mathbb{R}^p$, una funzione che soddisfa la proprietà di non negatività, identità e simmetria,

Il quadrato della distanza Euclidea (che si ottiene specificando come secondo argomento di `pdist`, la parola `'squaredeuclidean'`) non soddisfa la disuguaglianza triangolare².

Tabella 2.2: Matrice dei dati di partenza X . Numero di ordini e ammontare (in migliaia di Euro) effettuati da 4 clienti (A, \dots, D).

	Numero ordini	Ammontare
A	3	20
B	10	42
C	8	30
D	2	12

Consideriamo, ad esempio, la matrice dei dati riferita a 4 unità statistiche e due variabili (numero di ordini ed importo) riportata nella Tabella 2.2. La sintassi di seguito

²Nell'help della funzione `pdist`

<https://www.mathworks.com/help/stats/pdist.html#d123e650554>,

questo aspetto è sottolineato: `'squaredeuclidean'` *Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality).*

```

X=[3 20;10 42;8 30; 2 12];
% Calcolo della matrice dei quadrati delle
% distanze Euclidee.
D=squareform(pdist(X,'seuclidean'));
disp(D)

```

produce

0	533	125	65
533	0	148	964
125	148	0	360
65	964	360	0

È immediato osservare che

$$\begin{aligned}
 {}_2d_{12}^2 &> {}_2d_{13}^2 + {}_2d_{23}^2 \\
 533 &> 125 + 148
 \end{aligned}$$

in contrasto con la disuguaglianza triangolare dell'equazione (2.1).

Un'ulteriore categoria di indici è riportata dalla seguente:

Definizione: si dice indice di dissimilarità (*dissimilarity index*) tra i vettori x e $y \in \mathbb{R}^p$ ($ID(x, y)$) una funzione che gode delle proprietà di non negatività e simmetria e anche la seguente proprietà

$$x = y \Rightarrow d(x, y) = 0. \quad (2.3)$$

Di conseguenza, un indice di dissimilarità può essere uguale a 0 anche se $x \neq y$.

2.4 Lo spazio euclideo ponderato

Il ricercatore può decidere di attribuire un peso ad ogni variabile X_s della matrice dei dati che esprime il grado di importanza che vuole attribuire alla variabile medesima. Ad esempio, la ponderazione può servire ad attribuire maggiore importanza alle variabili che si ritengono essere più correlate con i fenomeni sottostanti in base ai quali si effettua la classificazione. In un'indagine sul benessere procapite sembra ragionevole attribuire maggiore importanza al reddito pro-capite rispetto al numero di telefoni cellulari per 100 abitanti. Quando la classificazione si basa su una pluralità di aspetti e per alcuni sono disponibili molti indicatori, mentre per altri le variabili sono in numero ridotto, la ponderazione può essere utilizzata per riequilibrare gli aspetti poco rappresentati. Similmente, si può pensare di dare un peso inferiore alle variabili misurate con minore grado di precisione.

Le distanze viste finora assumono implicitamente che le variabili in esame presentino tutto lo stesso ordine di grandezza e la stessa unità di misura (ad esempio numero di certificazioni ISO in due anni consecutivi, v. Tabella 2.1). In tutti i casi in cui questo non si verifica è necessario neutralizzare la diversa unità di misura dividendo ciascuna dimensione per una misura di variabilità.

Si definisce metrica di Minkowski ponderata la seguente espressione

$${}_{kw}d_{ij} = \left[\sum_{s=1}^p |x_{is} - x_{js}|^k w_s \right]^{1/k} \quad k > 0$$

dove w_s è il peso attribuito alla variabile s -esima. La distanza euclidea ponderata tra due vettori x_i e x_j è la norma euclidea ponderata della differenza tra essi

$${}_w d_{ij} = {}_w d_{ij} = \|x_i - x_j\|_w = \sqrt{(x_i - x_j)' W (x_i - x_j)} = \sqrt{\sum_{s=1}^p (x_{is} - x_{js})^2 w_s} \quad (2.4)$$

In generale la distanza Euclidea ponderata è una forma quadratica e la matrice della forma quadratica W è una matrice diagonale che contiene sulla diagonale principale i pesi assegnati alle diverse variabili: $W = \text{diag}(w_1, w_2, \dots, w_p)$. Se $W = \text{diag}(1/\text{var}(X_1), 1/\text{var}(X_2), \dots, 1/\text{var}(X_p))$ si ottiene la distanza euclidea calcolata sugli scostamenti standardizzati.

$${}_w d_{ij} = \|x_i - x_j\|_w = \sqrt{\frac{\sum_{j=1}^p (x_{is} - x_{js})^2}{\sigma_s^2}} = \sqrt{\sum_{j=1}^p (z_{is} - z_{js})^2} \quad (2.5)$$

Esercizio.

Partendo dalla matrice dei dati riportata nella Tabella 2.2, calcolare la matrice delle distanze eucldee sugli scostamenti standardizzati. Verificare che la chiamata a `pdist` con il secondo argomento di input `'seuclidean'` equivale ad effettuare il calcolo delle distanze eucldee sulla matrice Z .

```
% Inserimento dei dati
X=[3 20;10 42;8 30; 2 12];

% Calcolo della matrice delle distanze eucldee sugli
% scostamenti standardizzati

% Utilizzo l'opzione seuclidean
dist=pdist(X,'seuclidean');
```

```
% Standardizzo preliminarmente le variabili
distchk=pdist(zscore(X));
assert(max(abs(dist-distchk))<1e-12,"Errore di programmazione")
D=squareform(dist);
disp(D)
```

Il codice di cui sopra produce:

```
      0      2.4831      1.5071      0.6693
2.4831      0      1.0608      3.1061
1.5071      1.0608      0      2.0837
0.6693      3.1061      2.0837      0
```

Se il secondo argomento di input di `pdist` è uguale a `'seuclidean'` allora nel terzo argomento di input è possibile specificare il peso da assegnare alle variabili. Se il terzo argomento viene omesso `pdist(X,'seuclidean')` usa come pesi i reciproci delle varianze delle variabili originarie. Se si vogliono utilizzare stime robuste di σ come ad esempio il MAD riscalato, si può utilizzare una delle due sintassi che seguono

```
% MADS stime robuste di sigma
% norminv(0.75) è il valore esatto per il fattore di correzione 1.4826
MADs=mad(X,1)/norminv(0.75);
% Se il secondo argomento di input è seuclidean è possibile
% specificare come devono essere standardizzate le variabili
dist=pdist(X,'seuclidean',MADs);
% Nell'istruzione segue lavoro direttamente sugli scostamenti
```



```
% standardizzati robusti
distchk=pdist(zscoreFS(X));
assert(max(abs(dist-distchk))<1e-12,"Errore di programmazione nel " + ...
    "calcolo delle distanze Euclidee sulla matrice degli " + ...
    "scostamenti standardizzati robusti")
```

2.5 La distanza di Mahalanobis

Nelle distanze considerate finora non si tiene conto della correlazione tra le variabili originarie. L'introduzione anche di questo aspetto può far variare le valutazioni sulla diversità tra coppie di fenomeni. La distanza di Mahalanobis è una forma di distanza standardizzata, in cui si tiene conto non solo della diversa dispersione delle variabili, ma anche della loro correlazione.

Definizione: si dice distanza di Mahalanobis tra i vettori x_i e x_j l'espressione

$${}_M d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)} \quad (2.6)$$

dove S è la matrice di covarianze. Mentre nella distanza Euclidea i contorni di equidistanza in due dimensioni sono rappresentati da i punti che giacciono lungo una circonferenza, nella metrica di Mahalanobis i contorni di equidistanza sono rappresentati da ellissi. Nella sezione 1.1.7 (Figura 1.4) abbiamo visto che le curve di livello della distribuzione gaussiana multivariata, se le variabili sono correlate, sono ellissi. Un punto è tanto più probabile quanto più l'ellisse su cui giace è vicino al centro. Sembra quindi naturale definire una misura di distanza ("improbabilità") che cresce mano a mano che mi

sposto su ellissi più esterne.

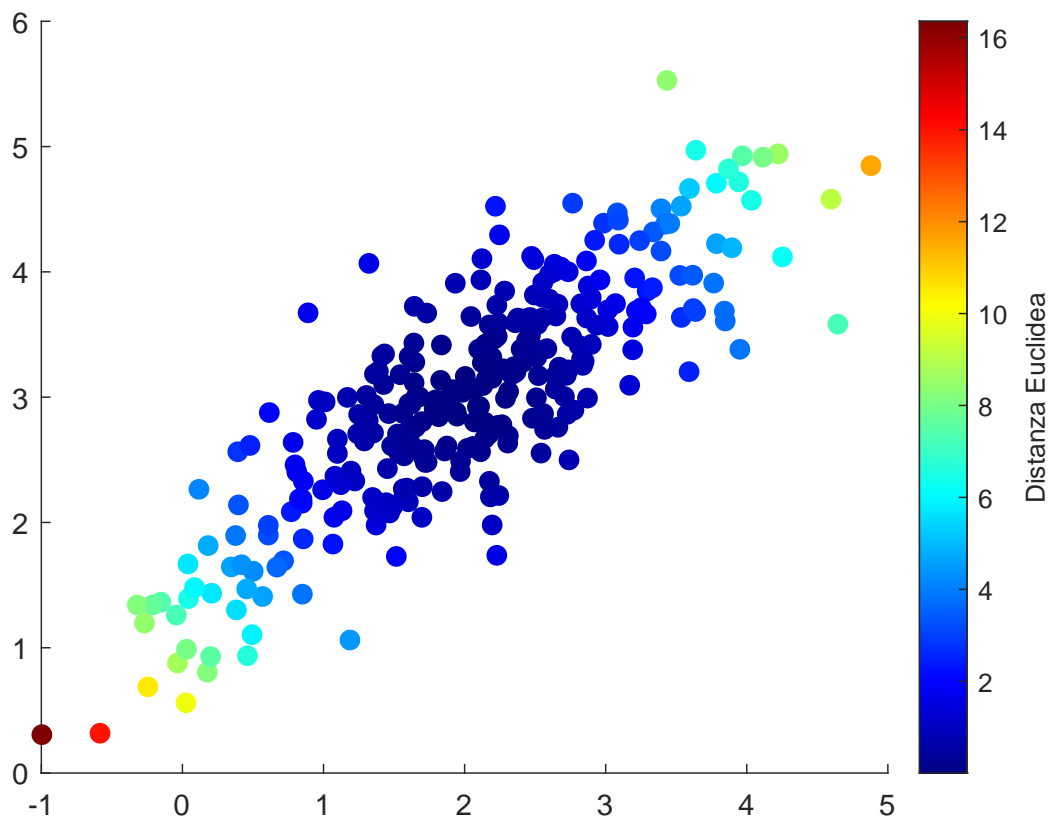


Figura 2.4: Diagramma di dispersione di 300 osservazioni generate da una distribuzione normale bivariata con componenti correlate positivamente. Il colore dei punti è proporzionale al valore del quadrato della distanza Euclidea. I punti più distanti dal centroide sono quelli che si trovano in alto a destra ed in basso a sinistra.

Finora abbiamo visto la distanza tra due generiche righe della matrice dei dati. È molto frequente dovere calcolare la distanza rispetto al centroide (media aritmetica delle osservazioni). In tal caso abbiamo

$${}_2d_{i\bar{x}} = \sqrt{(x_i - \bar{x})'(x_i - \bar{x})} \quad (2.7)$$

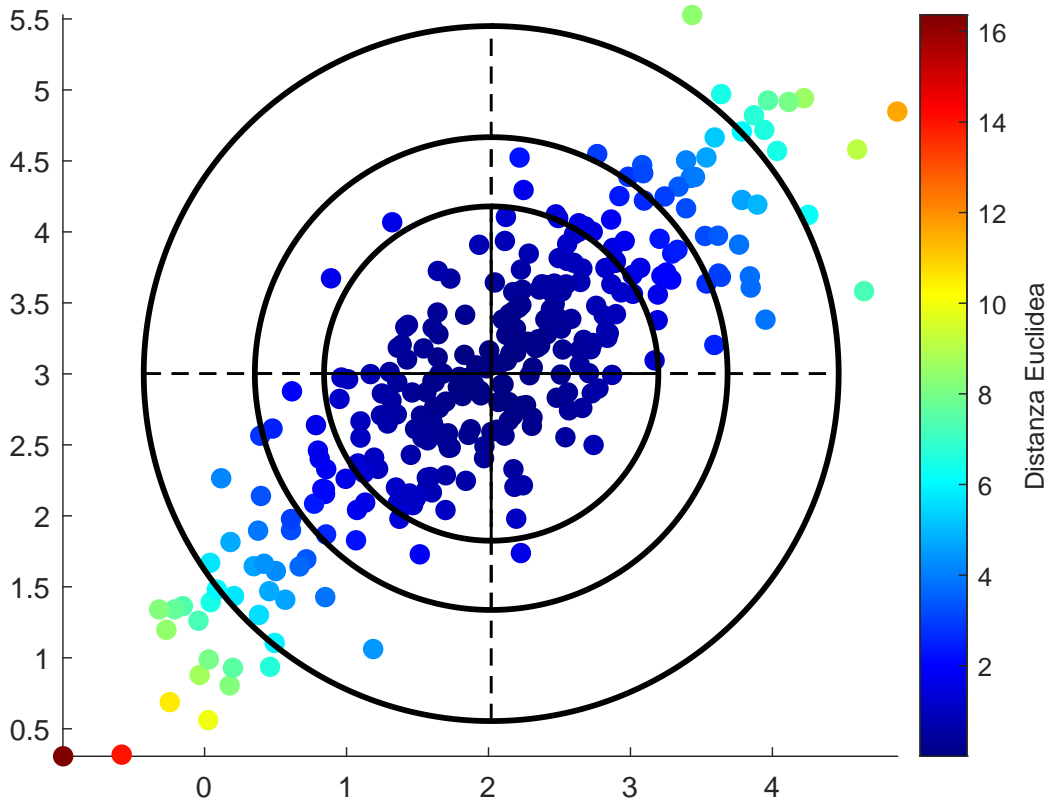


Figura 2.5: Aggiunta dei contorni di equidistanza al livello di confidenza 50% 75% e 95% ai punti della Figura 2.4. I contorni di equidistanza non tengono del fatto che le due variabili sono fortemente correlate.

$$_M d_{i\bar{x}} = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})} \quad (2.8)$$

Nell'esercizio che segue discutiamo queste due diverse distanze.

Esercizio.

Generare 300 realizzazioni casuali dalla distribuzione normale bivariata con parametri μ e Σ definiti come segue `mu=[2 3]; Sigma = [1.2 0.9; 0.9 0.8]`. Per garantire la replicabilità dei risultati utilizzare il seed `rng(1)`. Costruire un diagramma di dispersione in cui i punti hanno una size (ampiezza) di 50 ed il colore dei punti dipende dal valore della distanza Euclidea dal

centroide al quadrato. Aggiungere al grafico una barra di colore (`colorbar`) specificando una mappa di colore (`colormap`) di tipo `jet`. Creare un nuovo grafico in cui vengono aggiunti i contorni di equidistanza dal centroide utilizzando intervalli di confidenza al 50, 70 e 95 cento. Ripetere i due grafici precedenti con il quadrato della distanza di Mahalanobis.

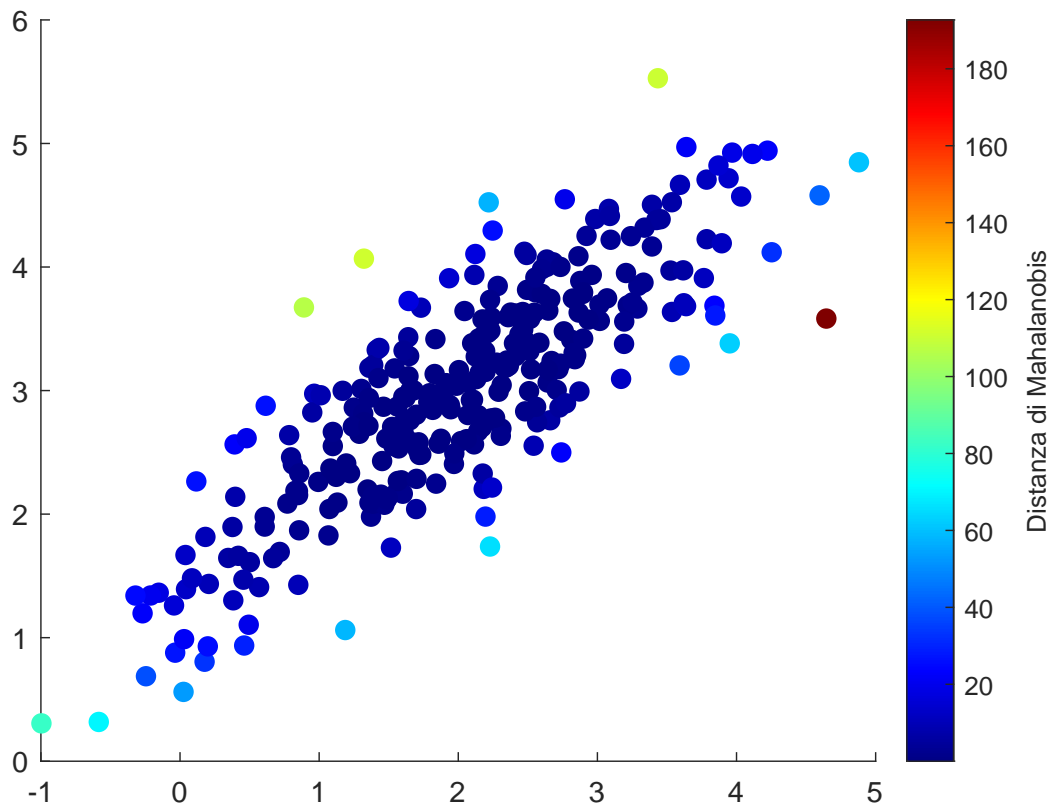


Figura 2.6: Diagramma di dispersione dei medesimi punti mostrati nella Figura 2.4 con colore proporzionale al valore del quadrato della distanza di Mahalanobis. I punti più distanti dal centroide sono quelli che si trovano lontano dalla nuvola dei punti nella direzione opposta a quella di massima variabilità.

Soluzione

```

%% Generazione della matrice 300x2 dalla distr. normale biv
rng(1)

```

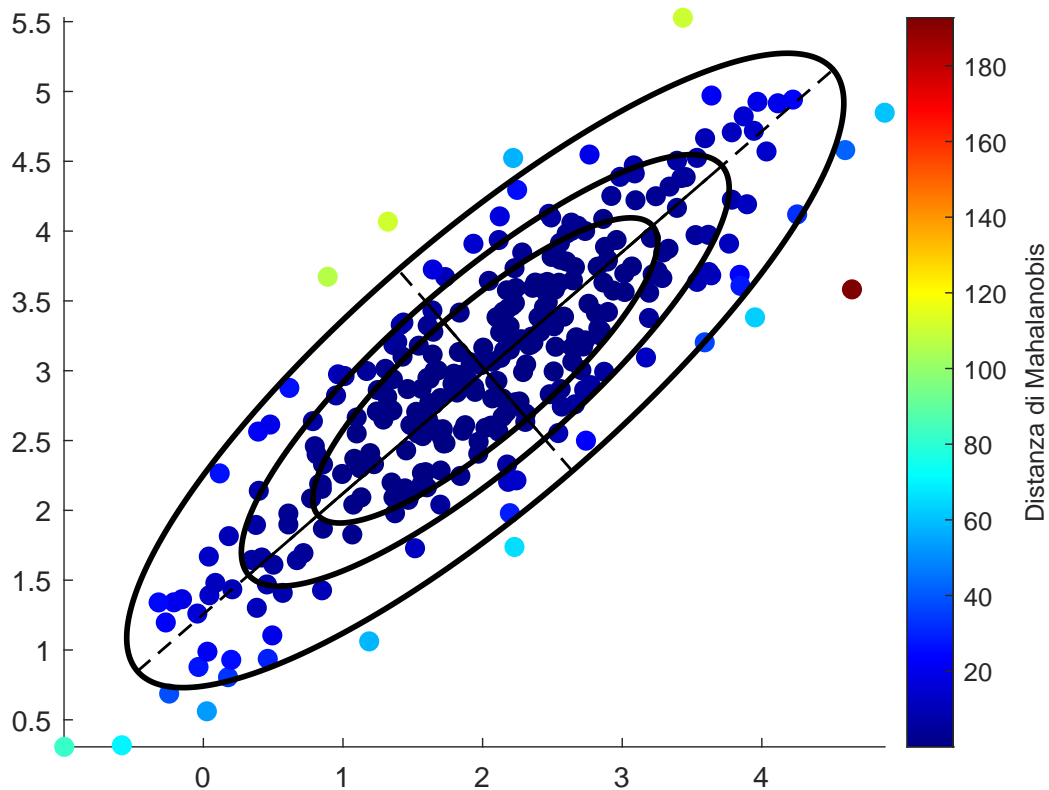


Figura 2.7: Aggiunta dei contorni di equidistanza al livello di confidenza 50% 75% e 95% ai punti della Figura 2.6. I contorni di equidistanza delle distanze di Mahalanobis tengono conto del fatto che le due variabili sono fortemente correlate.

```
n=300;

mu=[2;3];

covarianza=0.93;

R= [1.2 covarianza;covarianza 0.95];

% Genero i dati utilizzando la distribuzione normale bivariata
% con i valori di mu e Sigma specificati nel testo
X = mvnrnd(mu,R,n);

%% Calcolo della distanza Euclidea e di Mahalanobis dal centroide
```

```
% Calcolo del centroide
```

```
cent=mean(X);
```

```
% d2_Euclidean =vettore nx1 che contiene la distanza Euclidea al quadrato
% di ogni riga dal centroide
```

```
d2_Euclidean = sum((X-cent).^2,2);
```

```
% d2_mahal = =vettore nx1 che contiene la distanza di Mahalanobis al
% quadrato di ogni riga dal centroide
```

```
d2_mahal = mahal(X,X).^2;
```

```
% Implementazione utilizzando la funzione mahalFS dell'FSDA toolbox
```

```
ds_mahalFSDA=(mahalFS(X,cent,S)).^2;
```

```
%% Diagramma di dispersione con colore che dipende da dist Eucl
```

```
% Il quarto argomento di scatter è il colore. In questo caso il colore
% dipende dal valore della corrispondente distanza
```

```
scatter(X(:,1),X(:,2),50,d2_Euclidean,'o','filled')
```

```
hb = colorbar;
```

```
ylabel(hb,'Distanza Euclidea')
```

```
colormap jet
```

```
% Le osservazioni che presentano la più grande distanza Euclidea sono
% quelle in basso a sinistra e quelle in alto a destra.
```

```
%% Nuovo grafico con aggiunta dei contorni di equidistanza
```

```
figure
scatter(X(:,1),X(:,2),50,d2_Euclidean,'o','filled')
hb = colorbar;
ylabel(hb,'Distanza Euclidea')
colormap jet
prob=[0.50 0.75 0.95];
hold('on')
for j=1:3
    ellipse(cent,eye(2),prob(j))
end
axis equal
```

Questo codice produce le Figure 2.4 e 2.5. Per ottenere le Figure 2.6 e 2.7 è necessario, nel codice sopra, rimpiazzare l'istruzione

```
scatter(X(:,1),X(:,2),50,d2_Euclidean,'o','filled')

con

scatter(X(:,1),X(:,2),50,d2_mahal,'o','filled')
```

Nella distanza Euclidea i contorni di equidistanza sono dei cerchi, di conseguenza questa metrica deve essere utilizzata in presenza di variabili non correlate, ossia quando i punti in termini di scostamenti dalla media presentano la stessa proporzione in tutti e quattro i quadranti. Al contrario, se le variabili presentano una forte correlazione positiva, gli ellissi rappresentano i contorni corretti di equidistanza. In questa metrica i punti molto distanti sono quelli che si discostano in maniera marcata dalle direzioni di massima variabilità.

2.5.1 Proprietà della distanza di Mahalanobis

La distanza di Mahalanobis presenta importanti caratteristiche

1. Essa può interpretarsi come una distanza Euclidea ponderata definita nell'equazione (2.4) in cui la matrice dei pesi W è uguale all'inversa della matrice di covarianze ($W = S^{-1}$).
2. Essa non è funzione solo dei due vettori considerati x_i e x_j ma tiene conto del contesto in cui questi si collocano, cioè delle relazioni tra le variabili. Pertanto, anche se rimangono immutati i vettori x_i e x_j , la distanza di Mahalanobis può cambiare se muta la matrice di covarianze. Ad esempio la distanza di Mahalanobis tra due province può cambiare a seconda che si considerino solo le province della regione oppure tutte le province italiane.
3. Essa è invariante per qualsiasi trasformazione non singolare dei vettori riga nella matrice dei dati.

Dimostrazione

Utilizzando i vettori elementari e_i e e_j la distanza di Mahalanobis al quadrato può essere scritta:

$${}_M d_{ij}^2 = (e_i - e_j)' X S^{-1} X' (e_i - e_j)$$

Se la matrice X viene postmoltiplicata per una matrice A di dimensione

$p \times p$ che ammette l'inversa si ottiene

$$\begin{aligned}
 {}_M d_{ij}^2 &= (e_i - e_j)'(XA)(A'SA)^{-1}(XA)'(e_i - e_j) \\
 &= (e_i - e_j)'(XA) \left(A^{-1}S^{-1}(A')^{-1} \right) (XA)'(e_i - e_j) \\
 &= (e_i - e_j)'XAA^{-1}S^{-1}(A')^{-1}A'X'(e_i - e_j) \\
 &= (e_i - e_j)'XS^{-1}X'(e_i - e_j).
 \end{aligned}$$

La distanza di Mahalanobis, quindi, risulta invariante per trasformazioni di scala e/o di traslazione delle variabili originarie e per modifiche dei pesi delle stesse. In altri termini, le distanze di Mahalanobis sui dati originari oppure sui dati standardizzati non cambiano. È utile ricordare che la classe delle metriche di Minkowski non gode di questa proprietà.

4. Se le variabili sono tutte incorrelate tra loro, la distanza di Mahalanobis è uguale alla distanza euclidea calcolata sugli scostamenti standardizzati. In tal caso infatti

$$S^{-1} = \text{diag}(1/\text{var}(X_1), 1/\text{var}(X_2), \dots, 1/\text{var}(X_p))$$

e l'equazione (2.6) si riduce alla (2.5).

5. La somma delle distanze di Mahalanobis dal centroide al quadrato è

pari a $(n-1)p$.

Dimostrazione:

$$\begin{aligned}
 \sum_{i=1}^n M d_{i\bar{x}}^2 &= \sum_{i=1}^n (x_i - \bar{x})' S^{-1} (x_i - \bar{x}) & (2.9) \\
 &= \sum_{i=1}^n \text{tr}((x_i - \bar{x})' S^{-1} (x_i - \bar{x})) \\
 &= \sum_{i=1}^n \text{tr}(S^{-1} (x_i - \bar{x}) (x_i - \bar{x})') \\
 &= \text{tr}(S^{-1} \sum_{i=1}^n (x_i - \bar{x}) (x_i - \bar{x})') \\
 &= (n-1) \text{tr}(S^{-1} S) \\
 &= (n-1) \text{tr}(I_p) \\
 &= (n-1)p.
 \end{aligned}$$

2.6 La scala di misura delle distanze

I valori numerici assunti dalle distanze (o più in generale dagli indici di dissimilarità) possono essere interpretati in due maniere differenti:

1. si può pensare che tali valori numerici forniscano semplicemente l'ordinamento delle distanze tra coppie di vettori. In tal caso si parla di scala di misura ordinale delle distanze.
2. si può ritenere che i valori numerici rappresentino una misurazione della distanza. Essendo fissato univocamente lo zero del sistema di misura (condizione di identità nella definizione) le distanze si interpretano co-

me espresse su scala di rapporti.

Se si adotta l'interpretazione su scala di rapporti è lecito affermare che la distanza tra una determinata coppia è doppia rispetto a quella di un'altra coppia. Al contrario con la prima interpretazione si afferma semplicemente che una coppia di unità è più (meno) distante rispetto ad un'altra coppia.

Occorre osservare che anche se tutte le p variabili della matrice dei dati sono espresse su scala di rapporti, non è affatto pacifico che le distanze tra le unità debbano essere interpretate sulla stessa scala di misurazione in quanto esse rappresentano una funzione di sintesi delle differenze tra i vettori corrispondenti a due unità. In tale operazione di sintesi entrano in gioco scelte soggettive (standardizzazione delle variabili oppure tipo di standardizzazione) per cui la misura della diversità che si ottiene dipende da esse. Può quindi essere preferibile l'interpretazione delle distanze su scala semplicemente ordinale. Con questa interpretazione la matrice delle distanze Euclidee e la matrice dei quadrati di tali distanze sono ritenute equivalenti, poiché danno origine alla medesima graduatoria della dissimilarità tra coppie di elementi. Nell'esempio dell'esercizio precedente il coefficiente di cograduazione di Spearman calcolato come

```
corr([d2_Euclidean d2_mahal], 'type', 'Spearman')
```

risulta uguale a 0.8224. Le graduatorie quindi, in termini di distanze euclidee e di Mahalanobis non concordano, quindi non è possibile determinare univocamente la graduatoria delle distanze tra coppie di unità. Pertanto, anche interpretando le distanze su scala ordinale (la più debole) la valutazione della

dissimilarità tra le coppie di elementi non rimane invariata se si considera il tipo di distanza utilizzato per misurarla.

2.7 Gli indici di similarità

Definizione: dato un insieme finito di elementi $u_i \in U$, si dice indice di similarità un'applicazione $S(u_i, u_j) = S_{ij}$ da $U \times U$ in R^1 che soddisfa le seguenti condizioni:

1. non negatività:

$$S_{ij} \geq 0 \quad \text{per ogni} \quad u_i, u_j \in U$$

2. normalizzazione

$$S_{ii} = 1 \quad \text{per ogni} \quad u_i \in U$$

3. simmetria

$$S_{ij} = S_{ji} \quad \text{per ogni} \quad u_i, u_j \in U \quad (2.10)$$

Si noti che l'indice di similarità è definito con riferimento agli elementi di un insieme (unità statistiche) anziché ai vettori in \mathbb{R}^p come accadeva per

le distanze. Inoltre, un indice di similarità può assumere solo valori nell'intervallo chiuso $[0, 1]$ mentre una distanza può presentare qualsiasi valore nel campo dei numeri reali non negativi. Il complemento ad 1 di un indice di similarità $(1 - S_{ij})$ è un indice di dissimilarità ID_{ij} introdotto nella sezione 2.3. Infatti in base alla definizione di indici di similarità, ID_{ij} risulta sempre non negativo e simmetrico e la condizione di normalizzazione di S_{ij} implica che $ID_{ii} = 0$ (che equivale al vincolo (2.3) degli indici di dissimilarità).

2.7.1 Indici di similarità per fenomeni dicotomici

Una prima applicazione degli indici di similarità è quella riguardante n unità statistiche e p fenomeni tutti dicotomici. Ad esempio:

- per n esercizi alberghieri si rileva la presenza o l'assenza di p servizi accessori (piscina, salone per conferenze, sauna, televisore nelle camere, ...).
- per n consumatori si considerano p prodotti con modalità: acquistato oppure non acquistato;
- per n studenti universitari si rilevano p esami considerando per ciascuno di essi solo le modalità esame superato e esame non ancora superato.

L'obiettivo degli indici di similarità è quello di valutare la rassomiglianza tra le coppie di unità statistiche con riferimento ai caratteri considerati. Un

fenomeno dicotomico può sempre essere codificato in forma di una variabile indicatrice booleana che assume i valori 1=presenza=`true` oppure 0=assenza=`false`. In alcuni casi la presenza o assenza si riferisce direttamente al fenomeno (ad es. prodotto acquistato o meno). In altri casi, il carattere presenta due modalità contrapposte (ad esempio maschio o femmina) ed il ricercatore ne sceglie una e ne considera la presenza o assenza (ad es. si codificano i maschi con 1 e le femmine con 0).

Con riferimento a due unità statistiche u_i e u_j , i valori assunti dalle p variabili indicatrici, corrispondenti ai p caratteri dicotomici di partenza, possono essere classificati come riportato nella tabella che segue:

$u_i \backslash u_j$	1	0	Tot.
1	a	b	$a + b$
0	c	d	$c + d$
Tot.	$a + c$	$b + d$	p

dove:

a indica la frequenza dei fenomeni contemporaneamente presenti nelle due unità (co-presenze o *positive matches*).

d indica la frequenza dei fenomeni contemporaneamente assenti nelle due unità (co-assenze o *negative matches*).

b e c indicano la frequenza dei fenomeni presenti in un'unità ma non nell'altra.

Le frequenze b e c segnalano, quindi, gli aspetti di diversità tra le due unità statistiche considerate e devono essere trattate allo stesso modo per la condizione di simmetria (v. 2.10).

Le frequenze a e d indicano, invece, l'entità della rassomiglianza tra le due unità ma la loro importanza non è identica. La co-presenza di un carattere costituisce sempre un aspetto che concorre a definire la similarità. Al contrario, la co-assenza di un fenomeno in alcuni casi può risultare di scarso oppure nessun senso ai fini della valutazione della rassomiglianza tra due unità. Ad esempio, nell'analisi delle dotazioni di servizi in varie città, la co-presenza dell'aeroporto segnala sicuramente una rassomiglianza tra le due città (presumibilmente abbastanza grandi). Al contrario, la co-assenza dell'aeroporto si manifesta in città con caratteristiche molto diverse. Gli indici di similarità più noti sono i seguenti:

Definizione: si dice indice di similarità di Jaccard l'espressione:

$${}_JS_{ij} = \frac{a}{a + b + c} \quad (2.11)$$

Esso è pari al rapporto tra il numero di co-presenza ed il numero di caratteri considerati con l'esclusione di quelli che manifestano le co-assenza in u_i e u_j .

Definizione: si dice indice di similarità di Russel e Rao l'espressione:

$${}_{RR}S_{ij} = \frac{a}{a + b + c + d} = \frac{a}{p} \quad (2.12)$$

Esso è il rapporto tra il numero di co-presenze ed il numero totale di caratteri considerati. Gli indici ${}_JS_{ij}$ e ${}_{RR}S_{ij}$ al numeratore presentano solo le co-presenze. Un indice che dà uguale importanza alle co-presenze e co-assenze è il seguente:

Definizione: si dice indice di similarità di Sokal e Michener (oppure di

Hamming) l'espressione

$${}_SM S_{ij} = \frac{a + d}{a + b + c + d} = \frac{a + d}{p} \quad (2.13)$$

Esso è il rapporto tra il numero di caratteri che risultano uguali nelle due unità (co-presenze oppure co-assenze) ed il numero totale di caratteri considerati.

La Tabella 2.3 riporta le seguenti 8 caratteristiche dicotomiche di 6 alberghi:

RIS = presenza del ristorante;

CON= presenza della sala conferenze;

SUI= presenza di *suite*;

GAR= esistenza del garage;

HAN= presenza di stanze attrezzate per portatori di handicap;

CAT= categoria definita come il numero di stelle;

CAM = il numero di camere;

PRE = il prezzo massimo della camera doppia (PRE). Le prime 5 caratteristiche sono dicotomiche, al contrario le ultime 3 sono quantitative. Per quanto riguarda, ad esempio, gli alberghi C e D le 5 caratteristiche dicotomiche possono essere inserite nella seguente tabella

$C \setminus D$	1	0	Tot.
1	2	0	2
0	1	2	3
Tot.	3	2	5

segue immediatamente che

$${}_JS_{CD} = \frac{a}{a+b+c} = \frac{2}{2+0+1} = \frac{2}{3} \quad (2.14)$$

$${}_{RR}S_{CD} = \frac{a}{a+b+c+d} = \frac{2}{5} \quad (2.15)$$

$${}_{SM}S_{CD} = \frac{a+d}{p} = \frac{4}{5} \quad (2.16)$$

Esercizio. Partendo dai dati dicotomici contenuti nella Tabella 2.3, costruire la matrice degli indici di Jaccard e di corrispondenza semplice.

Tabella 2.3: Caratteristiche qualitative dicotomiche (prime 5 variabili) e quantattative (ultime tre variabili) di 6 alberghi

	RIS	CON	SUI	GAR	HAN	STE	CAM	PRE
A	1	1	1	1	1	5	169	450
B	1	1	1	1	0	4	94	410
C	0	1	0	1	0	4	48	330
D	1	1	0	1	0	4	66	330
E	0	0	0	0	0	3	32	190
F	0	0	0	1	0	3	33	165

Soluzione. Per calcolare le matrici richieste è sufficiente chiamare la funzione `pdist` con il secondo argomento rispettivamente pari a `'jaccard'` e `'hamming'` e trasformare le distanze in similarità.

```
%% Matrice di partenza
```

```
X=[1 1 1 1 1;
    1 1 1 1 0;
    0 1 0 1 0;
    1 1 0 1 0;
    0 0 0 0 0;
    0 0 0 1 0];
```

```
% rowlab = string array che contiene i nomi delle righe di X
rowlab=string(('A':'F')));

disp('Matrice di similarità di Jaccard')
SJ=1-squareform(pdist(X,"jaccard"));
StableJ=array2table(SJ,"RowNames",rowlab,'VariableNames',rowlab);
disp(StableJ)

disp('Matrice di similarità di Sokal Michener (Hamming)')
SSM=1-squareform(pdist(X,"hamming"));
SSMtable=array2table(SSM,"RowNames",rowlab,'VariableNames',rowlab);
disp(SSMtable)
```

Il codice di cui sopra produce l'output che segue:

Matrice di similarità di Jaccard

	A	B	C	D	E	F
	---	----	-----	-----	-	-----
A	1	0.8	0.4	0.6	0	0.2
B	0.8	1	0.5	0.75	0	0.25
C	0.4	0.5	1	0.66667	0	0.5
D	0.6	0.75	0.66667	1	0	0.33333
E	0	0	0	0	1	0
F	0.2	0.25	0.5	0.33333	0	1

Matrice di similarità di Sokal Michener (Hamming)

	A	B	C	D	E	F
	---	---	---	---	---	---
A	1	0.8	0.4	0.6	0	0.2
B	0.8	1	0.6	0.8	0.2	0.4
C	0.4	0.6	1	0.8	0.6	0.8
D	0.6	0.8	0.8	1	0.4	0.6
E	0	0.2	0.6	0.4	1	0.8
F	0.2	0.4	0.8	0.6	0.8	1

Per il calcolo degli indici di similarità di Russel e Rao si può procedere in due modi: tramite un semplice ciclo for (soluzione suggerita) oppure chiamando la funzione `pdist` con il secondo argomento di input che rappresenta un *function handle*. In questo caso è necessario creare una funzione di distanza personalizzata. Nel primo caso, il codice è riportato di seguito.

```
% S_RR = matrice degli indici di similarità di Russel Rao
% Viene inizializzata come una matrice di zeri di dimensione nxn
SRR=zeros(n,n);
for i=1:n
    % si confronta la riga i-esima della matrice X
    % con tutte le altre righe e si vanno a contare le coppie di 1
    % X(i,:) è un vettore riga di lunghezza p
    % X è una matrice di dimensione nxp
    % Tramite espansione implicita il vettore X(i,:) viene replicato
```

```

% n volte per renderlo conformabile con X
SRR(i,:)=sum(X(i,:)==1 & X==1,2)/p;
end
disp('Matrice di similarità di Russel e Rao')
SRRtable=array2table(SRR,"RowNames",rowlab,'VariableNames',rowlab);
disp(SRRtable)

```

Il codice di cui sopra produce il seguente output.

Matrice di similarità di Russel e Rao

	A	B	C	D	E	F
	---	---	---	---	-	---
A	1	0.8	0.4	0.6	0	0.2
B	0.8	0.8	0.4	0.6	0	0.2
C	0.4	0.4	0.4	0.4	0	0.2
D	0.6	0.6	0.4	0.6	0	0.2
E	0	0	0	0	0	0
F	0.2	0.2	0.2	0.2	0	0.2

Nel secondo modo `pdist` viene chiamata tramite l'istruzione `pdist(X,@simfun)`.

L'istruzione `@simfun` significa che la distanza viene definita tramite una funzione personalizzata costruita dall'utente che si chiama `simfun`. L'help della funzione `pdist` spiega che questa funzione deve essere costruita con due argomenti di input. Il primo è un vettore riga di lunghezza p ed il secondo è la matrice dei dati di dimensione $n \times p$. La nostra funzione personalizzata

deve essere in grado di restituire in output la distanza tra il vettore riga e ogni riga della matrice X .

```
% simfun = funzione personalizzata che calcola l'indice di similarità tra
% il vettore riga x1 e la matrice X
SRR1=pdist(X,@simfun);
disp(squareform(SRR1))

% simfun = funzione personalizzata che calcola l'indice di similarità tra
% il vettore riga x1 e la matrice X
function RR = simfun(x1,X)
% x1 è un vettore riga di lunghezza p
% X è una matrice di dimensione n x p
p=size(x1,2);
% RR = vettore colonna di lunghezza n, che contiene la similarità tra il vettore x1
% righe della matrice X
RR=sum(x1==1 & X==1,2)/p;
end
```

2.7.2 Indici di similarità in presenza di fenomeni misti

L'obiettivo di questa sezione è capire come calcolare la similarità quando si considerano contemporaneamente sia caratteri quantitativi e qualitativi (dicotomici o politomici). L'indice più conosciuto per la valutazione della prossimità in base congiuntamente a fenomeni qualitativi e quantitativi è l'indice proposto da Gower.

Definizione: si dice indice di similarità di Gower tra le unità u_i e u_j la seguente espressione:

$${}_G S_{ij} = \frac{\sum_{s=1}^p z_{ijs} w_{ijs}}{\sum_{s=1}^p w_{ijs}} \quad (2.17)$$

dove

$w_{ijs} = 1$ se è possibile il confronto tra le unità u_i e u_j per il fenomeno s -esimo.

$w_{ijs} = 0$ altrimenti.

Il confronto non è possibile se è mancante il dato del fenomeno s -esimo in almeno una delle due unità, oppure quando il fenomeno è dicotomico e si manifesta la co-assenza. Quindi, nel caso di co-assenza $w_{ijs} = 0$ vale 0 mentre le coppie 1-0 e 0-1 vale 1.

Il significato di z_{ijs} è diverso a seconda della scala di misura dei caratteri. In presenza di

- caratteri dicotomici
 - $z_{ijs} = 1$ se le unità u_i e u_j mostrano una co-presenza per il carattere s -esimo.
 - $z_{ijs} = 0$ altrimenti
- caratteri nominali con più di due modalità
 - $z_{ijs} = 1$ se le unità u_i e u_j presentano la stessa modalità per il carattere s -esimo.

– $z_{ijs} = 0$ altrimenti

- caratteri quantitativi

–

$$z_{ijs} = 1 - \frac{|x_{is} - x_{js}|}{\max(X_s) - \min(X_s)}$$

L'indice di similarità di Gower risulta uguale ad 1 se le unità i e j presentano valori identici per ciascuna delle variabili quantitative e modalità uguali per ognuno dei fenomeni qualitativi. Esso risulta uguale a zero nel caso di similarità nulla che corrisponde alla situazione in cui le unità i e j per ogni variabile quantitativa assumano l'una il valore massimo e l'altra il valore minimo e i caratteri qualitativi presentino modalità sempre diverse tra loro. Se tutti i fenomeni sono dicotomici, allora l'indice di Gower coincide con l'indice di similarità di Jaccard. Nel caso di sole variabili quantitative l'indice di Gower è uguale al complemento ad 1 della distanza media della città a blocchi calcolata sui valori rapportati al rispettivo campo di variazione. Ad esempio con riferimento agli alberghi C e D riportati nella Tabella 2.3, abbiamo che

$${}_G S_{CD} = \frac{0 + 1 + 0 + 1 + 0 + 1 - \frac{4-4}{5-3} + 1 - \frac{|48-66|}{169-32} + 1 - \frac{|330-330|}{450-165}}{6} = 0.8114$$

Si noti che il denominatore dell'indice di Gower è pari a 6 (e non a 8) in quanto le co-assenze dei fenomeni dicotomici non contano.

Esercizio: calcolare la matrice degli indici di similarità di Gower per i 6 alberghi riportati nella Tabella 2.3.

Soluzione. La funzione `Gowerindex` di FSDA consente di calcolare la matrice degli indici di Gower. In questa funzione uno degli argomenti opzionali di tipo `name/pairs` è il vettore denominato `l` di lunghezza p che consente di specificare la tipologia delle p -variabili.

$l(j)=1$ indica che la variabile j -esima è quantitativa;

$l(j)=2$, indica che la variabile j -esima è binaria.

$l(j)=3$, indica che la variabile j -esima è qualitativa polinomica, $j = 1, 2, \dots, p$.

Con riferimento ai dati riportati nella Tabella 2.3, `l` è definito come `l=[2 2 2 2 1 1 1]`.

Soluzione

```
X =[1 1 1 1 1 5 169 450 ;
    1 1 1 1 0 4 94 410;
    0 1 0 1 0 4 48 330;
    1 1 0 1 0 4 66 330;
    0 0 0 0 0 3 32 190;
    0 0 0 1 0 3 33 165];
l=[2*ones(1,5) ones(1,3)];
% rowlab = string array che contiene i nomi delle righe di X
rowlab=string(('A':'F'))';
% La funzione GowerIndex accetta anche un input di tipo table
Xtable=array2table(X,"RowNames",rowlab);
[IndSimG, IndSimGtable]=GowerIndex(Xtable,'l',l);
disp(IndSimGtable)
```

Il codice di cui sopra produce l'output che segue

	A	B	C	D	E	F
	-----	-----	-----	-----	-----	-----
A	1	0.72653	0.39947	0.54089	0.010965	0.12591
B	0.72653	1	0.62622	0.78785	0.18222	0.31359
C	0.39947	0.62622	1	0.81144	0.3784	0.56231
D	0.54089	0.78785	0.81144	1	0.29343	0.4467
E	0.010965	0.18222	0.3784	0.29343	1	0.72625
F	0.12591	0.31359	0.56231	0.4467	0.72625	1

Ad esempio ${}_G S_{AF} = 0.12591$ e segnala una modestissima similarità tra gli alberghi A e F. Le coppie di alberghi più simili risultano essere (A, B) e (E, F).

Gower ha dimostrato che la matrice degli indici di similarità che si ottiene è semidefinita positiva, purché non vi sia alcun *missing value* nella matrice dei dati di partenza.

Capitolo 3

La riduzione delle dimensioni

3.1 Analisi in componenti principali (PC): introduzione

Date p variabili quantitative X_1, X_2, \dots, X_p , l'obiettivo dell'analisi in componenti principali (PC, dall'inglese “Principal Components”) è quello di sostituire alle variabili originali, che possono essere correlate, un nuovo insieme di variabili, dette componenti principali che hanno le seguenti proprietà:

- sono incorrelate (ortogonali);
- sono elencate in ordine decrescente rispetto alla loro varianza.

La prima componente principale Y_1 è la combinazione lineare delle p variabili di partenza avente la massima varianza. La seconda componente principale Y_2 è la combinazione lineare delle p variabili con la varianza immediatamente inferiore alla varianza di Y_1 e ad essa incorrelata. ecc. fino

alla p -esima componente. Se le p variabili originali sono molto correlate, un numero $q < p$ tiene conto di una quota elevata di varianza totale per cui le prime q componenti forniscono una buona approssimazione di dimensione ridotta della struttura dei dati.

Lo scopo primario di questa tecnica è la riduzione di un numero elevato di variabili (rappresentanti altrettante caratteristiche del fenomeno analizzato) in alcune variabili latenti. Ciò avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale le variabili vengono ordinate in ordine decrescente di varianza: pertanto, la variabile con maggiore varianza viene proiettata sul primo asse, la seconda sul secondo asse e così via. La riduzione della complessità avviene limitandosi ad analizzare le principali (per varianza) tra le nuove variabili.

Ci sono 3 approcci per arrivare a determinare le componenti principali:

- Proiezioni di punti in un sottospazio
- Rappresentazione di una matrice di rango p con una matrice di rango ridotto
- Combinazione lineare delle variabili originarie

Nella sezione che segue illustriamo quest'ultimo approccio.

3.2 La prima PC come combinazione lineare delle variabili originarie

Obiettivo: data \tilde{X} (matrice degli scostamenti dalla media) oppure Z matrice degli scostamenti standardizzati, si cerca il vettore a dimensione $p \times 1$ in modo tale che sia massima $var(\tilde{X}a)$ oppure la $var(Za)$. Le soluzioni di questo problema di massimo sono infinite proporzionali, perché la combinazione lineare contiene un fattore di scala arbitrario. Si impone, quindi, il vincolo $a'a = 1$.

Proposizione: il vettore a che contiene i coefficienti della combinazione lineare delle variabili originarie che massimizza la varianza è il primo autovettore associato al primo autovalore della matrice di covarianze S (se si parte da \tilde{X} oppure è il primo ed primo autovettore associato al primo autovalore della matrice di correlazione R (se si parte da Z).

Dimostrazione.

Senza perdita di generalità supponiamo di partire dalla matrice Z (degli scostamenti standardizzati) e di voler massimizzare

$$\max_a var(y) = \max_a var(Za)$$

Le soluzioni di questo problema di massimo sono infinite proporzionali perché la combinazione lineare contiene un fattore di scala arbitrario. Si impone il vincolo che il vettore dei coefficienti abbia norma unitaria:

$$||a||^2 = a'a = 1$$

L'espressione della varianza di y è

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

In questo contesto \bar{y} è uguale a zero, in quanto

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{y'1_{n \times 1}}{n} = \frac{(Za)'1_{n \times 1}}{n} = \frac{a'Z'1_{n \times 1}}{n}$$

Tenendo presente che le colonne della matrice Z hanno somma pari a zero il vettore $p \times 1$, $Z'1_{n \times 1}$ è pari a

$$Z'1_{n \times 1} = \begin{pmatrix} \sum_{i=1}^n z_{i1} \\ \sum_{i=1}^n z_{i2} \\ \dots \\ \sum_{i=1}^n z_{ip} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix} = 0_{p \times 1}$$

Di conseguenza

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{a'0_{p \times 1}}{n} = 0$$

Possiamo quindi scrivere

$$\begin{aligned} var(y) &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n y_i^2}{n-1} = \frac{y'y}{n-1} \\ var(y) &= \frac{(Za)'Za}{n-1} = \frac{a'Z'Za}{n-1} = a' \frac{Z'Z}{n-1} a = a'Ra \end{aligned}$$

Si tratta quindi di massimizzare la forma quadratica $a'Ra$ dove R è la matrice di correlazione. Dato che la matrice di correlazione è semidefinita positiva la forma quadratica (ossia lo scalare) $a'Ra$ assume solo valori positivi. Il

nostro obiettivo, quindi, è quello di cercare il vettore a in modo tale che il numero $a'Ra$ sia il più grande possibile. Dato che $R = V\Lambda V'$ con $V'V = I_p$ (scomposizione spettrale, v. sezione 1.4).

$$var(y) = a'Ra = a'V\Lambda V'a = (V'a)'\Lambda(V'a) = b'\Lambda b$$

dove $b = V'a$. Se ricordiamo che la matrice V è ortogonale abbiamo che $a = Vb$. Il vincolo $a'a = 1$ impone quindi che $(Vb)'Vb = b'V'Vb = b'b = 1$. Quindi

$$var(y) = b'\Lambda b$$

con il vincolo $b'b = 1$. L'espressione $b'\Lambda b$ non è altro che una somma dei quadrati ponderati:

$$var(y) = b'\Lambda b = \sum_{i=1}^p b_i^2 \lambda_i \leq \sum_{i=1}^p b_i^2 \lambda_{\max}$$

dove λ_{\max} è il più grande autovalore della matrice R . Tenendo presente che $b'b = 1$

$$var(y) \leq \sum_{i=1}^p b_i^2 \lambda_{\max} = \lambda_{\max} \sum_{i=1}^p b_i^2 = \lambda_{\max} b'b = \lambda_{\max}$$

In conclusione, il numero più grande che può assumere la forma quadratica $a'Ra$ è uguale all'autovalore più grande della matrice R . Affinché la forma quadratica raggiunga il suo massimo ($var(y) = \lambda_{\max}$) come deve essere scelto il vettore a ? Supponiamo senza perdita di generalità che gli autovalori siano ordinati in senso non decrescente $\lambda_{\max} = \lambda_1 \geq \lambda_2, \dots, \geq \lambda_p$.

$$var(y) = a'V\Lambda V'a = \lambda_1 = b'\Lambda b$$

implica che b (affinché la forma quadratica estragga l'elemento 1,1 di Λ) deve essere uguale a $e'_1 = (1, 0, 0, \dots, 0)$. ossia che

$$V'a = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Questo accade quando $a = v_1$ ossia quando a è l'autovettore associato a λ_1 .

In tal caso infatti

$$V'v_1 = \begin{pmatrix} v'_1 \\ v'_2 \\ \vdots \\ v'_p \end{pmatrix} v_1 = \begin{pmatrix} v'_1 v_1 \\ v'_2 v_1 \\ \vdots \\ v'_p v_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

dato che $V'V = I_p$ e $V = (v_1, v_2, \dots, v_p)$. In conclusione, il vettore a che contiene i coefficienti della combinazione lineare delle variabili originarie che massimizza la varianza è il primo autovettore associato al primo autovalore della matrice R , se si parte dalla matrice degli scostamenti standardizzati, oppure della matrice S se si parte dalla matrice degli scostamenti dalla media \tilde{X} .

3.3 Le prime k PC come combinazioni lineari delle variabili originarie

In generale il nostro obiettivo è trovare k vettori a_1, a_2, \dots, a_k ciascuno combinazione lineare delle variabili originarie (senza perdita di generalità supponiamo di partire dalla matrice degli scostamenti standardizzati)

$$y_1 = Za_1, \quad y_2 = Za_2, \quad \dots, \quad y_r = Za_k$$

in termini matriciali

$$Y_{n \times k} = (y_1, y_2, \dots, y_k) = Z_{n \times p} A_{p \times k}$$

dove A è la matrice di dimensione $p \times k$, $A = (a_1, a_2, \dots, a_k)$ tale per cui la somma delle varianze della matrice Y sia la più grande possibile. In simboli, l'obiettivo è trovare la matrice A che massimizza la seguente espressione:

$$\max_A \text{tr cov}(Y) = \max_A \text{tr cov}(ZA)$$

con il vincolo che $A'A = I_r$. Questa espressione è massimizzata quando $A = V = (v_1, v_2, \dots, v_k)$ ossia quando A contiene i primi k autovettori associati ai primi k autovalori della matrice R . Se invece che partire dalla matrice Z fossimo partiti dalla matrice \tilde{X} , allora l'espressione precedente sarebbe stata massimizzata quando A contiene i primi k autovettori associati ai primi k autovalori della matrice S .

Dimostrazione

$$\max_A \text{tr} \text{cov}(Y) = \max_A \text{tr} \text{cov}(ZA) = \max_A \text{tr} \frac{A'Z'ZA}{n-1} = \max_A \text{tr} A'RA$$

con il vincolo che $A'A = I_k$. La matrice $A'RA$ può essere scritta come:

$$A'RA = \begin{pmatrix} a'_1Ra_1 & a'_1Ra_2 & \dots & a'_1Ra_k \\ a'_2Ra_1 & a'_2Ra_2 & \dots & a'_2Ra_k \\ \vdots & \vdots & \ddots & \vdots \\ a'_kRa_1 & a'_kRa_2 & \dots & a'_kRa_k \end{pmatrix}$$

La traccia di questa matrice è

$$\text{tr}(A'RA) = \sum_{i=1}^k a'_iRa_i$$

Dato che a'_iRa_i è massima quando $a_i = v_i$, ($i = 1, 2, \dots, k$) $\text{tr}(\text{cov}(A)) = \text{tr}(A'RA)$ è massima quando la matrice A contiene i primi k autovettori della matrice R se si parte da Z (oppure i primi k autovettori della matrice S se si parte da \tilde{X}).

3.3.1 Relazione tra autovalori traccia e determinante

L'obiettivo di questa sezione è quello di dimostrare la relazione che sussiste tra la traccia (somma degli elementi sulla diagonale principale) di una matrice quadrata e simmetrica, il suo determinante ed i suoi autovalori.

$$\text{tr}(A) = \text{tr}(V\Lambda V') = \text{tr}(\Lambda V'V) = \text{tr}(\Lambda) = \sum_{i=1}^p \lambda_i$$

di conseguenza, la somma degli autovalori di una matrice simmetrica è pari alla somma degli elementi sulla sua diagonale principale.

Osservazione: nel caso in cui la generica matrice A , su cui si vuole fare la scomposizione spettrale, sia la matrice di covarianze S di ordine p , si ha che

$$tr(S) = \sum_{i=1}^p var(X_i) = tr(\Lambda) = \sum_{i=1}^p \lambda_i$$

la somma degli autovalori, quindi, non è altro che la somma delle varianze delle variabili originarie. Di conseguenza, il rapporto:

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i} = \frac{\sum_{i=1}^k \lambda_i}{tr(S)} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p var(X_i)} \quad (3.1)$$

indica la quota di varianza delle p variabili originarie che viene spiegata dalle prime k componenti principali. Si noti che nel caso in cui si operi sulla matrice di correlazione

$$\sum_{i=1}^p var(X_i) = \sum_{i=1}^p \lambda_i = p$$

Se invece dell'operatore traccia applichiamo l'operatore determinante si ottiene che:

$$|A| = |V\Lambda V'| = |\Lambda V'V| = |\Lambda| = \prod_{i=1}^p \lambda_i$$

ossia il prodotto degli autovalori è esattamente uguale al determinante della matrice A . Se la matrice A è la matrice di covarianze allora, dato che $|S|$ può interpretarsi come una varianza generalizzata, l'estrazione delle prime k componenti principali massimizza la varianza generalizzata che si può estrarre

considerando k combinazioni lineari delle variabili originarie.

Osservazione: in questa sezione per dimostrare la relazione tra traccia, determinante e autovalori abbiamo utilizzato la scomposizione spettrale (cioè abbiamo supposto che la matrice fosse simmetrica). In realtà la relazione vista in questa sezione vale per qualsiasi matrice A non necessariamente simmetrica purché i suoi autovalori siano reali. È facile verificare che nell'esempio visto nella sezione 1.2

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix}$$

la somma degli autovalori $1 + 4 = 5$ coincide con $tr(A) = 2 + 3 = 5$ e che il prodotto degli autovalori $1 \times 4 = 4$ coincide con il determinante $det(A) = |A| = 6 - 2 = 4$.

3.4 La scomposizione in valori singolari (svd)

Qualsiasi matrice X di dimensioni $n \times p$ di rango r (con $r \leq \min(n, p)$) può essere scomposta come

$$X_{n \times p} = U_{n \times r} \Gamma_{r \times r} V'_{r \times p} \quad (3.2)$$

dove

$U = (u_1, u_2, \dots, u_r)$ è una matrice ortogonale $U'U = I_r$ che contiene gli autovettori di XX' .

$V = (v_1, v_2, \dots, v_r)$ è una matrice ortogonale $V'V = I_r$ che contiene gli

autovettori di $X'X$.

Γ è una matrice diagonale che contiene sulla diagonale principale i valori singolari (ossia le radici quadrate) degli r autovalori non nulli $\lambda_1, \lambda_2, \dots, \lambda_r$ delle matrici $X'X$ oppure XX' .

$$\Gamma = \begin{pmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \gamma_2 & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \gamma_r \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_r} \end{pmatrix} = \Lambda^{0.5}$$

Si osservi che le colonne della matrice $Y = XV$ sono tra loro incorrelate, ossia la matrice di covarianze della matrice XV è diagonale:

$$\text{cov}(Y) = \text{cov}(XV) = (U\Gamma)'(U\Gamma)/(n-1) = \Lambda/(n-1)$$

Questo significa che nella scomposizione in valori singolari data una determinata matrice di partenza (diciamo A) e due generici vettori di norma unitaria (diciamo x e y), dal punto di vista geometrico la svd si ha quando i due vettori Ax e Ay sono tra loro ortogonali (perpendicolari). Questo caso è esemplificato alla p. web

<https://blogs.mathworks.com/cleve/2013/08/05/eigshow-week-3>

Nel nostro esempio precedente, x e y sono due generiche colonne della matrice degli autovettori V (ossia due vettori ortogonali, ossia due vettori che formano sempre nello spazio a due dimensioni un angolo di 90 gradi). I vettori Ax e Ay sono due nuovi vettori non necessariamente ortogonali. Se i due nuovi vettori Ax e Ay vengono scelti come ortogonali allora si ha la svd.

Nel nostro contesto la matrice A non è altro che \tilde{X} (oppure la matrice Z), i vettori Ax e Ay non solo altro che due generiche componenti principali $\tilde{X}v_i$ e $\tilde{X}v_j$ (nuovi vettori ortogonali) e identificano rispettivamente due nuovi assi. I vettori $\tilde{X}v_i$ e $\tilde{X}v_j$ non sono altro che multipli delle colonne di U in quanto $Xv_i = U_i\gamma_i$ e $Xv_j = U_j\gamma_j$.

<https://blogs.mathworks.com/cleve/2016/08/08/bug-report-revives-interest-in-svd-option-of-eigshow/>

In presenza di due sole dimensioni, le prime due componenti principali identificano rispettivamente gli assi maggiore e minore dell'ellisse che può essere sovrapposto al diagramma di dispersione. La lunghezza dei semiassi dell'ellisse è esattamente uguale ai valori singolari della matrice \tilde{X} (oppure Z). In termini algebrici la lunghezza di $\tilde{X}v_i/\sqrt{n-1}$, ossia la norma Euclidea della i -esima componente principale, ossia la radice quadrata della varianza della i -esima componente principale, coincide con il semiasse i -esimo dell'ellisse:

$$\sqrt{\text{var}(y_i)} = \sqrt{\text{var}(\tilde{X}v_i)} = \sqrt{(\tilde{X}v_i)'(\tilde{X}v_i)/(n-1)} \quad (3.3)$$

$$= \sqrt{v_i'(\tilde{X}'\tilde{X})/(n-1)v_i} = \sqrt{v_i'Sv_i} = \sqrt{v_i'V\Lambda V'v_i} \quad (3.4)$$

$$= \sqrt{e_i'\Lambda e_i} = \sqrt{\lambda_i} = \gamma_i \quad (3.5)$$

dove e_i è un vettore colonna composto da elementi tutti uguali a 0 eccetto per la posizione i -esima dove il valore che assume è 1. $v_i'V = v_i'(v_1, v_2, \dots, v_p) = e_i'$ in quanto $v_i'v_j = 0$ per $i \neq j$ e $v_i'v_i = 1$ per $i = j$. Dato che, per definizione della prima componente principale la norma precedente viene massimizzata,

la prima componente principale nel caso di due dimensioni identifica l'asse maggiore dell'ellisse. Dato che la seconda componente principale è ortogonale alla prima, essa identifica l'asse minore dell'ellisse.

Un modo alternativo di scrivere la s.v.d. è il seguente:

$$X = \sum_{i=1}^r \gamma_i u_i v_i'$$

che mostra come qualsiasi generica matrice X di rango r può essere scritta come somma di r matrici di dimensioni $n \times p$ ciascuna di rango 1

La matrice $\gamma_i u_i v_i'$ è di rango 1, in quanto per le proprietà del rango (v. sezione 1.1.15):

$$\begin{aligned} \text{rank}(u_i v_i') &= \text{rank}((u_i v_i')' u_i v_i') = \text{rank}(v_i u_i' u_i v_i') \\ &= \text{rank}(v_i v_i') \leq \min(\text{rank}(v_i), \text{rank}(v_i')) = \min(1, 1) = 1 \end{aligned}$$

La funzione `[U, Gamma, V] = svd(A, 'econ')` consente di calcolare nei 3 argomenti di output `U`, `Gamma` e `V` le 3 matrici viste sopra. È facile verificare in Matlab che $A = U * \text{Gamma} * V'$.

3.5 Le prime k PC come migliore rappresentazione di rango k delle variabili originali

Il nostro obiettivo è quello di sostituire al posto della matrice $X_{n \times p}$ di partenza di rango r (il cui generico elemento è x_{ij}) una nuova matrice $\hat{X}_{n \times p}$ (il cui generico elemento è \hat{x}_{ij}) di rango ridotto, in modo tale che sia minima la somma dei quadrati delle differenze

$$\min_{\hat{x}_{ij}} (x_{ij} - \hat{x}_{ij})^2$$

Dato che la somma dei quadrati degli elementi di una matrice si può scrivere come:

$$\sum_{i=1}^n \sum_{j=1}^p x_{ij}^2 = \text{tr}(X'X)$$

si ottiene che

$$\min_{\hat{x}_{ij}} (x_{ij} - \hat{x}_{ij})^2 = \text{tr}[(X - \hat{X})'(X - \hat{X})]$$

Se per esempio cerchiamo come matrice \hat{X} una matrice di rango 2 abbiamo che:

$$X - \hat{X} = \sum_{i=1}^r \gamma_i u_i v_i' - \sum_{i=1}^2 \gamma_i u_i v_i' = \sum_{i=3}^r \gamma_i u_i v_i'$$

3.5. LE PRIME K PC COME MIGLIORE RAPPRESENTAZIONE DI RANGO K DELLE VARIABILI

$$\begin{aligned}
 tr \left((X - \hat{X})'(X - \hat{X}) \right) &= tr \left(\sum_{i=3}^r \gamma_i u_i v_i' \sum_{i=3}^r \gamma_i u_i v_i' \right) \\
 &= tr \left(\sum_{i=3}^r \sum_{j=3}^r v_i \gamma_i u_i' u_j \gamma_j v_j' \right) \\
 &= tr \left(\sum_{i=3}^r \sum_{j=3}^r v_i \gamma_i \gamma_j v_j' \right) \\
 &= tr \left(\sum_{i=3}^r \sum_{j=3}^r \gamma_i \gamma_j v_j' v_i \right) \\
 &= tr \left(\sum_{i=3}^r \gamma_i^2 v_i' v_i \right) \\
 &= tr \left(\sum_{i=3}^r \gamma_i^2 \right) \\
 &= \sum_{i=3}^r \gamma_i^2 \\
 &= \sum_{i=3}^r \lambda_i
 \end{aligned}$$

In altri termini, la miglior rappresentazione di rango 2 della matrice originaria (ossia quella che minimizza la somma dei quadrati dei residui) si ottiene quando si prendono i primi due autovettori associati ai primi due autovalori della matrice $X'X$ (XX') oppure della matrice $\tilde{X}'\tilde{X}$, ($\tilde{X}\tilde{X}'$).

Se si dividono le precedenti equazioni per $(n - 1)$ si ha che

$$tr \left((X - \hat{X})'(X - \hat{X}) \right) / (n - 1)$$

è pari $\sum_{i=3}^r \lambda_i$ dove in questo caso i λ_i sono gli autovalori della matrice

$\tilde{X}'\tilde{X}/(n-1)$ ossia gli autovalori della matrice di covarianze.

Se come matrice X si sceglie la matrice \tilde{X} la scomposizione svd può essere scritta come

$$\begin{aligned}\tilde{X} &= \sqrt{n-1}U \frac{\Gamma^*}{\sqrt{n-1}}V' \\ \tilde{X}/\sqrt{n-1} &= U\Gamma V' \\ \tilde{X} &= \sqrt{n-1}U\Gamma V'\end{aligned}$$

dove $\Gamma = \frac{\Gamma^*}{\sqrt{n-1}}$ è la matrice che contiene le radici quadrate degli autovalori della matrice $\tilde{X}'\tilde{X}/(n-1) = S$ ossia le radici quadrate degli autovalori della matrice di covarianze.

La matrice $\sqrt{n-1}U\Gamma = \tilde{X}V = Y$ contiene i valori della prime r componenti principali.

La matrice V contiene gli autovettori della matrice $\tilde{X}'\tilde{X}/(n-1) = S$ ossia gli autovettori della matrice di varianze e covarianze S (oppure gli autovettori della matrice di correlazione se si parte dalla matrice Z anziché dalla matrice \tilde{X}).

Dato che:

$$tr(\tilde{X}'\tilde{X})/(n-1) = tr(S) = \sum_{i=1}^p var(X_i)$$

possiamo riassumere affermando che, se al posto della matrice originaria di dimensioni $n \times p$ di rango p si sostituisce la miglior approssimazione di rango k , la quota non spiegata di varianza delle variabili originarie è esattamente

3.5. LE PRIME K PC COME MIGLIORE RAPPRESENTAZIONE DI RANGO K DELLE VARIABILI

uguale a:

$$\frac{\sum_{i=k+1}^p \lambda_i}{\sum_{i=1}^p \text{var}(X_i)}$$

Si noti che questo risultato è esattamente uguale identico a quello ottenuto nella sezione 3.3, equazione (3.1) in cui le prime k componenti principali erano state introdotte come combinazione lineare delle variabili originarie.

Si noti che

$$\text{cov}(\sqrt{n-1}U\Gamma) = \text{cov}(\tilde{X}V) = \text{cov}(Y) \quad (3.6)$$

$$= (\tilde{X}V)'(\tilde{X}V)/(n-1) = V'\tilde{X}'\tilde{X}/(n-1)V \quad (3.7)$$

$$= V'SV = V'V\Lambda V'V = \Lambda. \quad (3.8)$$

La matrice di varianze e covarianze della matrice delle componenti principali è diagonale (ossia le componenti principali sono tra loro incorrelate) e la varianza della i -esima componente principale è pari all' i -esimo autovalore λ_i della matrice di covarianze.

$$\begin{aligned} \text{cov}(\sqrt{n-1}U) &= \text{cov}(ZV\Gamma^{-1}) \\ &= (ZV\Gamma^{-1})'(ZV\Gamma^{-1})/(n-1) \\ &= \Gamma^{-1}V'(Z'Z/(n-1))V\Gamma^{-1} \\ &= \Gamma^{-1}V'RV\Gamma^{-1} = \Gamma^{-1}V'V\Lambda V'V\Gamma^{-1} \\ &= \Gamma^{-1}\Lambda\Gamma^{-1} = \Gamma^{-1}\Gamma\Gamma\Gamma^{-1} = I_p \end{aligned} \quad (3.9)$$

La matrice $\sqrt{n-1}U$, quindi, contiene i valori delle componenti principali

standardizzate (ossia con varianza unitaria).

3.6 PC come proiezione ortogonale dei punti in un sottospazio di dimensione ridotta

La matrice dei dati dimensione $n \times p$ descrive un insieme di vettori (riga o colonna) che individuano una nuvola di punti (rispettivamente unità o variabili). L'insieme delle distanze a due a due tra tutti i punti individua la forma della nuvola dei punti. Tale forma caratterizza la natura e l'intensità delle relazioni tra i punti e quindi rivela la struttura dell'informazione contenuta nei dati. Un modo per rendere visibile questa informazione nello spazio a p dimensioni è quello di proiettarla su delle rette o dei piani a due dimensioni. Il problema è dunque quello di cercare la retta oppure il sottospazio migliore, ossia quello che consente la minor perdita di informazione possibile. Fra i criteri di adattamento di un sottospazio ad una nuvola di punti si ricorre al criterio dei minimi quadrati, che consiste nel ricercare la retta dalla quale risulti minima la somma dei quadrati delle distanze $x - \hat{x}$ (v. Figura 1.10). Senza perdita di generalità supponiamo di lavorare con i vettori \tilde{x}_i che compongono la matrice \tilde{X} degli scostamenti dalla media. In simboli l'obiettivo è

$$\min_{\hat{x}_i} \sum_{i=1}^n \|\tilde{x}_i - \hat{x}_i\|^2$$

Essendo la proiezione della nuvola dei punti ortogonale alla retta cercata (v. ad esempio la Figura 1.9), per il teorema di Pitagora applicato ad ogni punto

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

\tilde{x}_i si ha che

$$\min_{\hat{x}_i} \sum_{i=1}^n \|\tilde{x}_i - \hat{x}_i\|^2 = \min_{\hat{x}_i} \left(\sum_{i=1}^n \|\tilde{x}_i\|^2 - \sum_{i=1}^n \|\hat{x}_i\|^2 \right)$$

La quantità

$$\sum_{i=1}^n \|\tilde{x}_i\|^2 = \sum_{i=1}^n \sum_{j=1}^p \tilde{x}_{ij}^2 = (n-1) \sum_{j=1}^p \text{var}(X_j) = (n-1) \text{tr}(S)$$

(v. la sezione sulla traccia) è uguale alla devianza totale ed è indipendente dai vettori \hat{x}_i cercati. Minimizzare $\sum_{i=1}^n \|\tilde{x}_i - \hat{x}_i\|^2$ equivale, quindi, a massimizzare la seguente espressione:

$$\max_{\hat{x}_i} \sum_{i=1}^n \|\hat{x}_i\|^2$$

Tenendo presente la formula della lunghezza della proiezione ed il fatto che il vettore a su cui si proietta è a norma unitaria, allora l'obiettivo è

$$\max_a \sum_{i=1}^n \|\tilde{x}_i' a\|^2 = a' \sum_{i=1}^n \tilde{x}_i \tilde{x}_i' a = \max_a (n-1) a' S a \quad a' a = 1$$

L'equazione da massimizzare è quindi (a meno del fattore $(n-1)$) la stessa che si ottiene nel caso in cui si voglia trovare la combinazione lineare con la massima varianza (v. sezione 3.2). Il massimo della funzione si ottiene quando a è l'autovettore associato al più grande autovalore della matrice di covarianze S .

3.6.1 Retta di regressione e retta associata alla prima componente principale

L'obiettivo di questa sezione è quello di sottolineare la differenza tra la retta che individua la prima componente principale e la retta di regressione. Con la regressione semplice e multipla si minimizza la somma dei quadrati delle distanze tra y_i e \hat{y}_i . Nell'analisi in componenti principali la retta migliore è quella che minimizza le distanze dai punti ortogonalmente rispetto ad essa. Consideriamo a titolo di esempio i dati riportati nella tabella 1.2, la costruzione del vettore delle medie aritmetiche \bar{x}' delle due variabili e della matrice di covarianze S attraverso il codice che segue:

```
% Dati caricati in formato table
Xtable=readtable("SpesaFatt.xlsx");

% X matrice dei dati
X=Xtable{:, :};

n=size(X,1);

X1=X(:,1);
X2=X(:,2);

% Vettore riga delle medie aritmetiche
meaX=mean(X);

% Matrice degli scostamenti dalla media
Xtilde=X-meaX;

% S = matrice di covarianze
S=Xtilde'*Xtilde/(n-1);
```

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

Nella sezione 1.1.7 abbiamo visto che l'equazione $x'Sx = c^2$ definisce un'ellisse. In questa sezione spieghiamo perché l'asse maggiore di questa ellisse passa attraverso la direzione di massima variabilità. Utilizzando il codice che segue possiamo sovrapporre al diagramma di dispersione gli ellissi di confidenza al 50, 75, 90 e 99 per cento: (v. Figura 3.1):

```
plot(X1,X2,'o','LineWidth',3)
xlabel('X1=Spesa pubblicitaria (mln €)')
ylabel('X2=Fatturato (mln €)')
hold('on')
confLevEllipses=[0.5 0.75 0.90 0.99];
for i=1:length(confLevEllipses)
    ellipse(meaX,S,confLevEllipses(i));
end
axis equal
```

A questo punto si trovano gli autovalori (ordinati in senso non decrescente) ed i corrispondenti autovettori della matrice S

```
% Autovalori ed autovettori di S
[Vini,Lambdaini]=eig(S);
[~,ord]=sort(diag(Lambdaini),'descend');
% Lambda contiene sulla diagonale, gli
% autovalori ordinati in senso decrescente
Lambda=Lambdaini(ord,ord);
% V contiene i corrispondenti autovettori
V=Vini(:,ord);
```

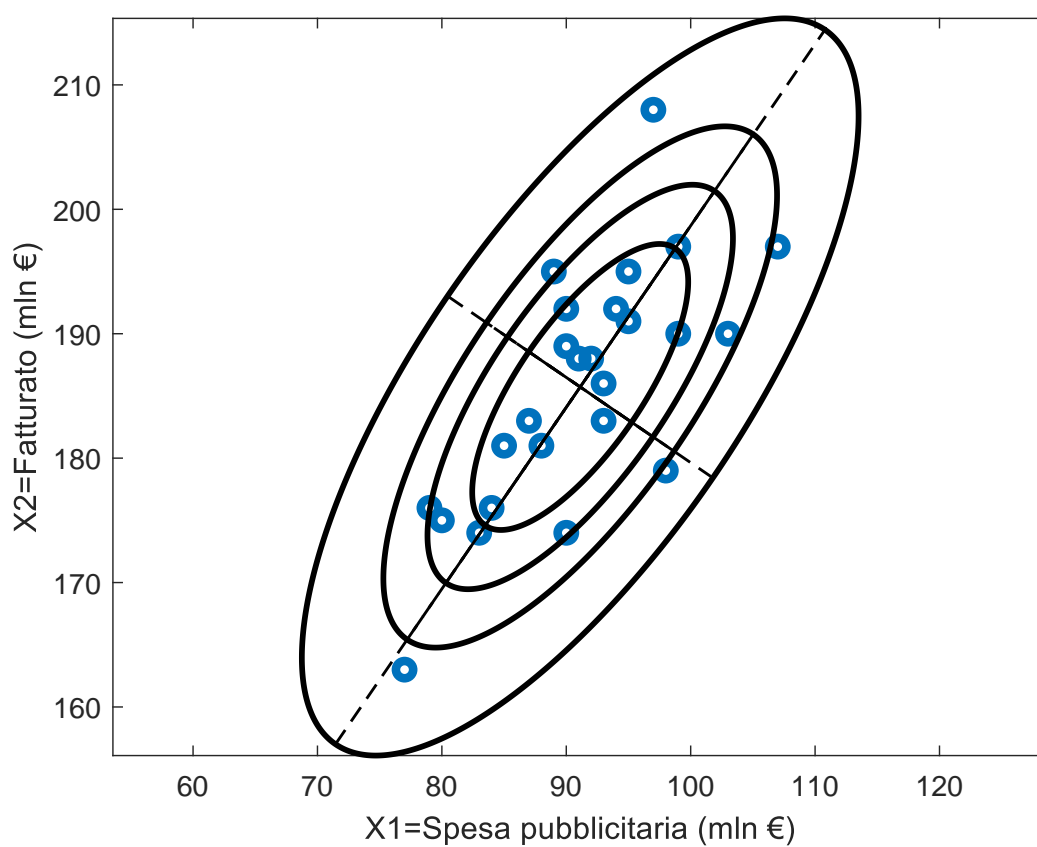


Figura 3.1: Diagramma di dispersione e ellissi di confidenza al 50, 75, 90, 99 per cento.

$$\Lambda = \begin{pmatrix} 131.5183 & 0 \\ 0 & 18.1350 \end{pmatrix} \quad V = (v_1 \ v_2) = \begin{pmatrix} 0.5652 & -0.8249 \\ 0.8249 & 0.5652 \end{pmatrix} \quad (3.10)$$

L'equazione della prima componente principale si ottiene ponendo uguale

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

a zero il valore della seconda componente principale.

$$\begin{aligned}v_2' \begin{pmatrix} X_1 - \bar{x}_1 \\ X_2 - \bar{x}_2 \end{pmatrix} &= 0 \\v_{12}(X_1 - \bar{x}_1) + v_{22}(X_2 - \bar{x}_2) &= 0 \\X_2 &= -\frac{v_{12}}{v_{22}}(X_1 - \bar{x}_1) + \bar{x}_2 \\X_2 &= b_{princ}X_1 + a_{princ}\end{aligned}$$

dove $b_{princ} = -\frac{v_{12}}{v_{22}}$ e $a_{princ} = \bar{x}_2 - b_{princ}\bar{x}_1$.

Il codice per trovare i coefficienti b_{princ} e a_{princ} ed aggiungere la retta principale è riportato di seguito (v. Figura 3.2)

```
figure
plot(X1,X2,'o')
xlabel('X1=Spesa pubblicitaria (mln €)')
ylabel('X2=Fatturato (mln €)')
bprinc=-V(1,2)/V(2,2);
aprinc=meaX(2)-bprinc*meaX(1);
% Viene aggiunta la retta principale
refline(bprinc,aprinc);
```

Il vettore $y_1 = \tilde{X}v_1$ contiene le coordinate dei punti nello spazio della prima componente principale (rispetto al primo autovettore). Si noti che y_1 ha dimensione $n \times 1$. Come abbiamo visto nell'equazione (1.12), le coordinate dei punti nello spazio originario a due dimensioni in termini di scostamenti dalla media ($\tilde{X}_1 \quad \tilde{X}_2$), sono date da y_1v_1' . Si noti che $y_1v_1' = \tilde{X}v_1v_1'$ ha dimensione $n \times 2$. Questa matrice è la miglior rappresentazione di rango 1

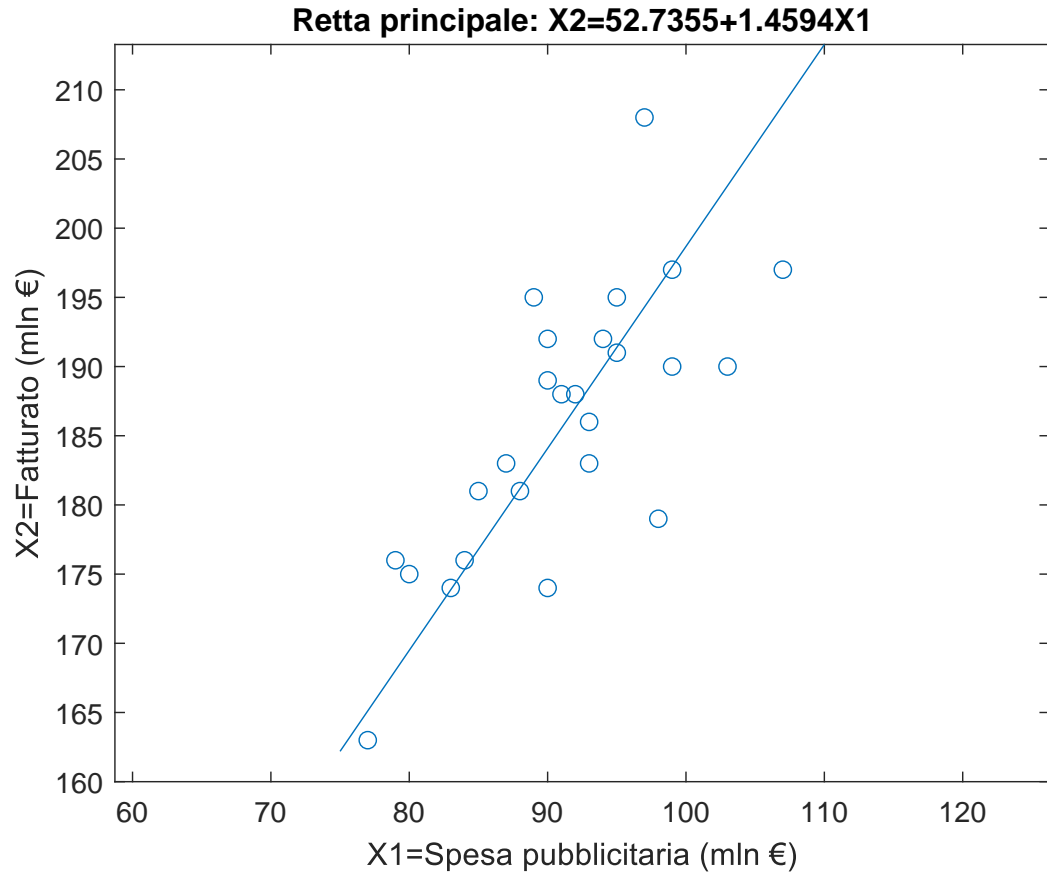


Figura 3.2: Diagramma di dispersione e retta principale (retta che passa per l'asse maggiore dell'ellisse) sovrapposta ai punti.

della matrice originaria \tilde{X} . Se si uniscono con segmenti i punti originali e le loro proiezioni sulla retta principale otteniamo la Figura 3.3.

Il codice per effettuare questa operazione è riportato di seguito

```
hold('on')

% y1 coordinate dei punti nello spazio della prima PC
% y1 è ad una sola dimensione
v1=V(:,1);
```

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE 2

```
y1=Xtilde*v1;

% I punti sono proiettati nello spazio originario
% a due dimensioni.
% Xtildehat sono le coordinate delle proiezioni ortogonali lungo la retta
% principale che passa per l'origine (punti in termini di scostamenti dalla
% media).
Xtildehat=y1.*v1';
% Le coordinate di Xtildehat vanno traslate tramite il vettore delle medie
% aritmetiche in modo tale da proiettare i punti lungo la retta principale
% che passa per il centroide (ossia il vettore delle medie aritmetiche)
Xhat=Xtildehat+meaX;

% Vengono aggiunte le linee che si riferiscono alle proiezioni
% ortogonali dei punti lungo la retta principale.
plot([Xhat(:,1) X1]',[Xhat(:,2) X2]','k')

% Con axis equal si usa la stessa lunghezza per le unità
% di misura lungo ciascun asse.
axis equal
```

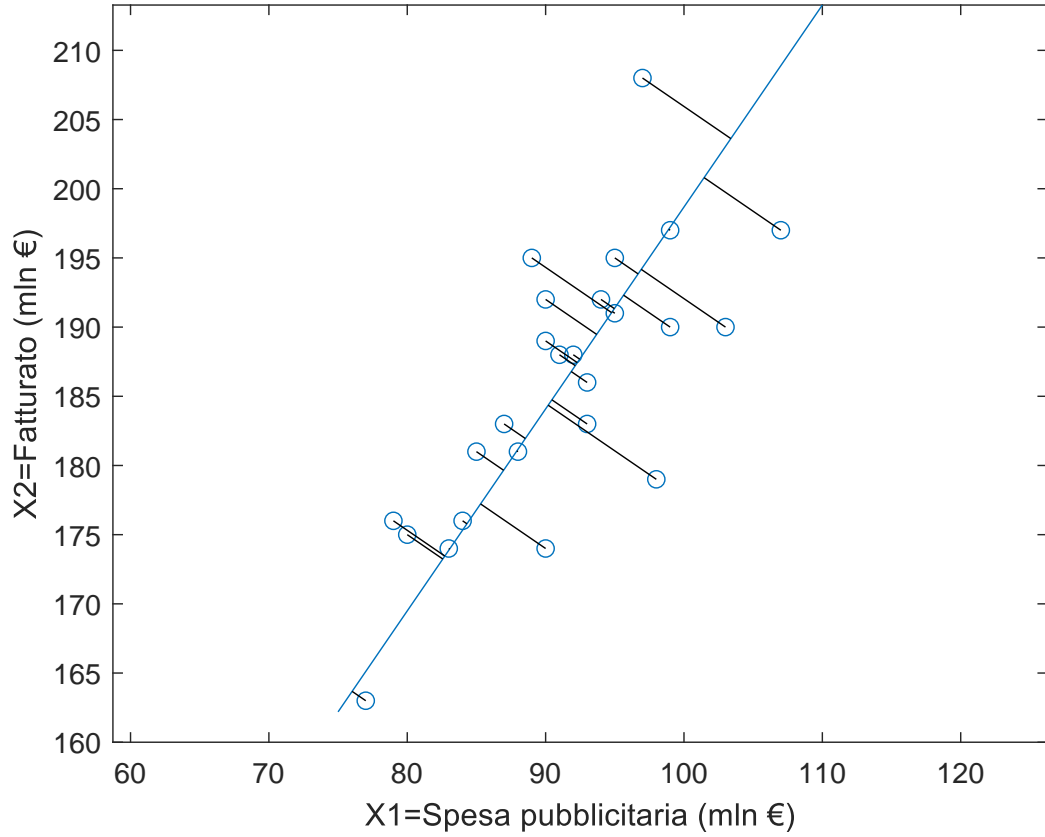


Figura 3.3: Diagramma di dispersione, retta principale e proiezione ortogonale dei punti lungo la retta principale che passa per i valori medi delle due variabili.

3.6.2 Ricostruzione della matrice originaria con una matrice di rango ridotto

La Tabella 1.2 conteneva i dati di partenza. Nella Tabella 3.1 abbiamo riportato sia i dati originari, sia la matrice $[\hat{X}_1 \quad \hat{X}_2]$, ossia la matrice $\tilde{X}v_1v_1' + 1_{n \times 1}\bar{x}'$ ossia le coordinate delle proiezioni dei punti sulla retta principale. La matrice $\tilde{X}v_1v_1'$ rappresenta la miglior approssimazione di rango

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

uno della matrice degli scarti dalla media¹. Le differenze tra X e \hat{X} (la sua

Tabella 3.1: Spesa pubblicitaria e fatturato per 25 aziende del settore tessile (dati in milioni di Euro) nelle prime due colonne. Nelle rimanenti due colonne la matrice $\hat{X} = \tilde{X}v_1v_1' + 1_{n \times 1}\bar{x}'$

Spesapub	Fatturato	\hat{X}_1	\hat{X}_2
95	191	94.82158282	191.1222502
89	195	94.76975272	191.0466071
88	181	87.92234721	181.0532071
93	183	90.45236338	184.7456222
84	176	84.31298371	175.7855458
97	208	103.3873185	203.6234555
90	189	92.29156665	187.4298351
99	197	98.8972259	197.0704201
92	188	92.46426987	187.6818856
90	192	93.69040539	189.4713609
98	179	90.18470207	184.3549857
87	183	88.53541497	181.9479447
90	174	85.29737296	177.2222061
99	190	95.63326885	192.3068599
91	188	92.14477847	187.2156061
77	163	76.01490936	163.674977
95	195	96.68670114	193.8442846
93	186	91.85120212	186.787148
85	181	86.963873	179.6543684
80	175	82.56873852	173.2399189
94	192	94.968371	191.3364792
83	174	83.06093315	173.958249
79	176	82.7155267	173.4541479
107	197	101.4531571	200.8006567
103	190	96.91123446	194.1719782

ricostruzione con una matrice di rango ridotto) sono i residui delle componenti principali. La funzione per calcolare direttamente i residui e ottenere la matrice \hat{X} si chiama `pcares`.

¹ $\text{rank}(\tilde{X}v_1v_1')$ risulta uguale ad 1

La somma dei quadrati delle differenze divisa per $n - 1$, tra X e la sua ricostruzione \hat{X}

```
sum((X-Xhat).^2,'all')/(n-1)
```

```
ans =
```

```
18.1350
```

è esattamente uguale al secondo autovalore della matrice S (v. equazione 3.10).

Dimostrazione

Per le proprietà della traccia la somma dei quadrati delle differenze si può scrivere come:

$$\begin{aligned}
 \text{tr}((\tilde{X} - \tilde{X}v_1v_1')'(\tilde{X} - \tilde{X}v_1v_1')) &= \\
 \text{tr}((\tilde{X}v_2v_2')'(\tilde{X}v_2v_2')) &= \\
 \text{tr}(v_2v_2'\tilde{X}'\tilde{X}v_2v_2') &= \\
 \text{tr}(v_2'\tilde{X}'\tilde{X}v_2) &= \\
 (n-1)\text{tr}(v_2'Sv_2) &= \\
 (n-1)\text{tr}(v_2'V\Lambda V'v_2) &= \\
 (n-1)\text{tr}(e_2'\Lambda e_2) &= \\
 (n-1)\lambda_2
 \end{aligned}$$

Con il simbolo e_2 si intende il vettore elementare $(0 \ 1)'$ (v. equazione 1.7).

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

Concludiamo questa sezione sottolineando le differenze tra la retta principale e la retta di regressione che si ottiene minimizzando la somma dei quadrati delle differenze tra i valori effettivi ed i valori teorici. Il codice che segue produce la Figura 3.4.

```
% Si cerca la retta che minimizza la somma dei quadrati dei residui tra i
% valori osservati della variabile dipendente (y_i) ed i valori adattati
%  $\hat{y}_i = a + bx_i$ , con  $i=1, 2, \dots, n$ 
figure
% X2= Fatturato= variabile dipendente
% X1= Spesa pubblicitaria = variabile esplicativa
% fitlm è la routine MATLAB per calcolare tutte le statistiche relative al
% modello di regressione
out=fitlm(X1,X2);
% Dalla struct out vengono estratti i parametri a
% e b della retta di regressione ed i valori adattati (yhat)
areg=out.Coefficients.Estimate(1);
breg=out.Coefficients.Estimate(2);
% yhat = valori adattati
yhat=out.Fitted;
% scatter dei valori originari
plot(X1,X2,'o')
hold('on')

% il comando lsline aggiunge la retta di regressione
% al diagramma di dispersione
```

```

hOLS=lsline;
% Colore rosso per la retta di regressione
hOLS.Color='r';
% Vengono aggiunte le distanze verticali dei punti dalla retta
plot([X1 X1]', [yhat X2]', 'k')
xlabel('X1=Spesa pubblicitaria (mln €)')
ylabel('X2=Fatturato (mln €)')
title(['Retta di regressione: X2=' num2str(areg) '+' num2str(breg) 'X1'])

axis equal

```

3.6.3 Componenti principali come rotazione degli assi cartesiani

Nella sezione precedente abbiamo visto che limitarsi ad analizzare la prima componente significa proiettare i punti nella direzione di massima variabilità (asse maggiore dell'ellissoide di confidenza). Nel caso in cui si considerano tutte le componenti principali non c'è alcuna riduzione delle dimensioni ma semplicemente una rotazione degli assi. Il luogo dei punti (X_1, X_2) tale per cui

$$(X_1 - \bar{x}_1 \ X_2 - \bar{x}_2)S^{-1}(X_1 - \bar{x}_1 \ X_2 - \bar{x}_2)' = 1 \quad (3.11)$$

è rappresentato nella Figura 3.5. Per tracciare l'ellisse è stata utilizzata l'istruzione `E11=ellipse(meaX,S,1)`. Ci possiamo chiedere qual è l'area, il perimetro e la lunghezza del semiasse principale di questo ellisse.

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

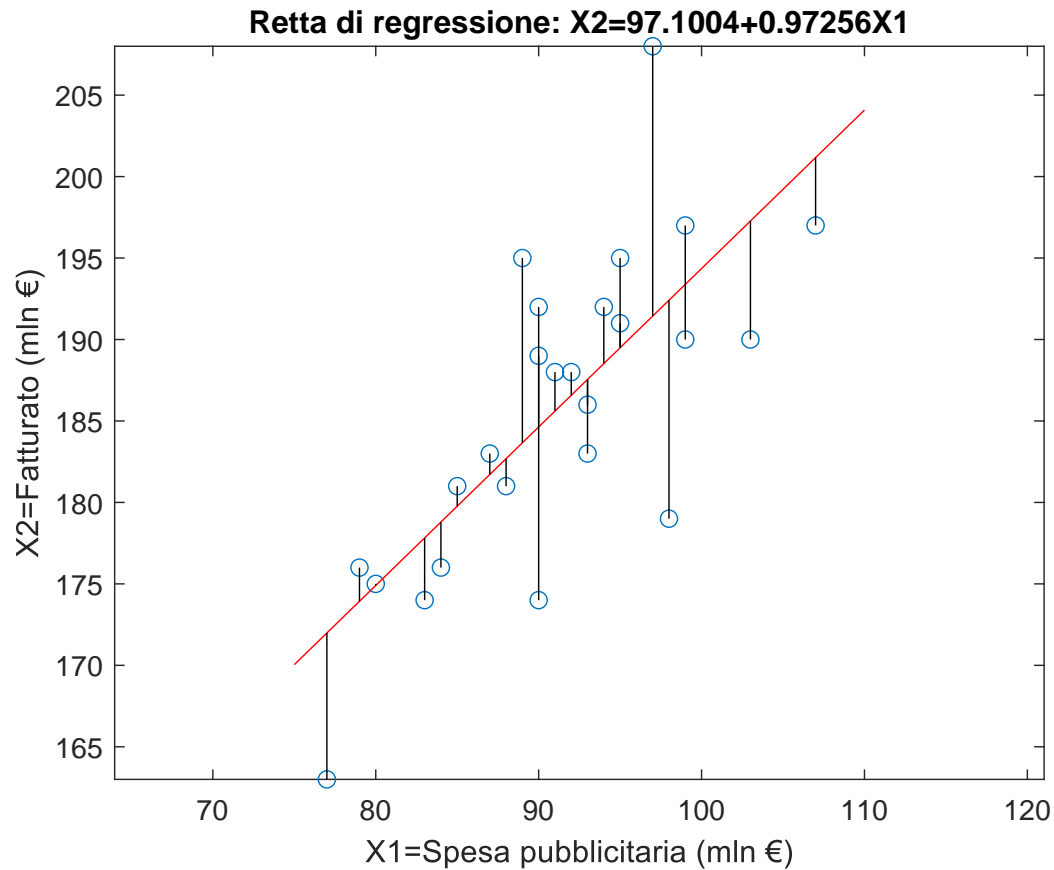


Figura 3.4: Diagramma di dispersione, retta di regressione dei minimi quadrati e proiezioni verticali dei punti lungo la retta di regressione che passa per i valori medi delle due variabili.

In questa sezione mostriamo inizialmente come effettuare questi calcoli in maniera numerica e poi mostriamo che queste quantità sono tutte funzioni dei valori singolari della matrice di covarianze S .

Per calcolare in maniera numerica la lunghezza del semiasse occorre fare l'intersezione tra la retta principale e i punti dell'ellisse secondo le linee mostrate nella sezione 1.5. Il codice è riportato di seguito:

```
% Trasformo le coordinate dell'ellisse in un poligono
```

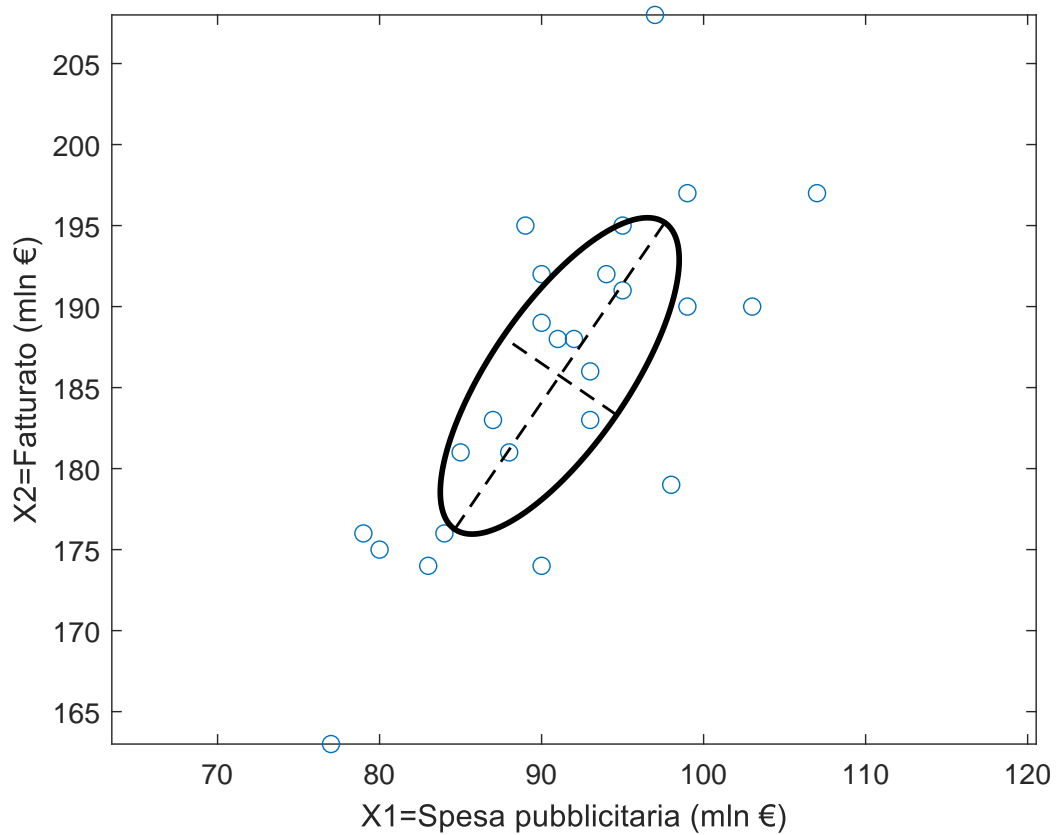


Figura 3.5: Diagramma di dispersione e luogo dei punti $(X_1 - \bar{x}_1 \ X_2 - \bar{x}_2)' S^{-1} (X_1 - \bar{x}_1 \ X_2 - \bar{x}_2)' = 1$

```

ellp=polyshape(Ell);
% Coordinate x e y della retta principale
xcoo=(min(X1):0.01:max(X1))';
ycoo=aprinc+bprinc*xcoo;
% Trovo l'intersezione tra la retta principale e l'ellisse
[in,out]=intersect(ellp,[xcoo ycoo]);
disp("Lunghezza del semiasse principale dell'ellisse")
disp("ottenuta in maniera numerica")
% in(end,:) contiene le coordinate dell'ultimo punto della retta

```

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

```
% che si trova dentro l'ellisse
% lunSemiAxis contiene la lunghezza del semiasse
lunSemiAxis=norm(in(end,:)-meaX);
% La norma poteva anche esser calcolata come segue
% sqrt(sum((in(end,:)-meaX).^2))
disp(lunSemiAxis)
```

Questo codice produce

```
Lunghezza del semiasse principale dell'ellisse
ottenuta in maniera numerica
11.4681
```

È facile controllare che il numero 11.4681 non è altro che la radice quadrata dell'elemento 1,1 della matrice Λ ottenuto nell'espressione (3.10).

Come già sottolineato nella sezione 3.4, le componenti principali identificano rispettivamente gli assi maggiore e minore dell'ellisse che può essere sovrapposto al diagramma di dispersione (v. sezione 1.5. La lunghezza del semiasse maggiore dell'ellisse è esattamente uguale al primo valore singolare della matrice \tilde{X} (oppure della matrice Z se fossimo partiti dagli scostamenti standardizzati). In termini algebrici la lunghezza di $\tilde{X}v_1/\sqrt{n-1}$, ossia la norma Euclidea della prima componente principale, ossia la radice quadrata della varianza della prima componente principale, coincide con il semiasse maggiore dell'ellisse².

²Nella sezione 3.6.3 vedremo il dettaglio computazionale di questa affermazione.

L'area ed il perimetro dell'ellisse in maniera numerica possono essere calcolati applicando le funzioni `area` e `perimeter` all'output della funzione `polyshape` (v. sezione 1.5). Il codice

```
disp(['Area dell''ellisse: ' num2str(area(ellp))])
disp(['Perimetro dell''ellisse: ' num2str(perimeter(ellp))])
```

produce come output

```
Area dell'ellisse: 153.4244
Perimetro dell'ellisse: 52.0965
```

La formula per il calcolo dell'area dell'ellisse è il prodotto delle lunghezze dei due semiassi moltiplicata per π^3 . L'istruzione

```
pi*prod(diag(sqrt(Lambda)))
```

produce 153.4270 un valore molto vicino a quello ottenuto in maniera numerica.

Per quanto riguarda il perimetro dell'ellisse, non esiste una forma chiusa ma una serie di formule approssimate. Applicando, ad esempio la formula approssimata dovuta al matematico Indiano Ramanujan⁴:

$$\text{Perimetro ellisse} \approx \pi \left(3(\gamma_1 + \gamma_2) - \sqrt{(3\gamma_1 + \gamma_2)(\gamma_1 + 3\gamma_2)} \right)$$

otteniamo 52.0376 un valore molto vicino a quello ottenuto in maniera numerica in precedenza.

³Si veda esempio, <https://it.wikipedia.org/wiki/Ellisse>.

⁴Il codice MATLAB è `gamma1=sqrt(Lambda(1,1)); gamma2=sqrt(Lambda(2,2)); disp(pi*(3*(gamma1+gamma2)-sqrt((3*gamma1+gamma2)*(gamma1+3*gamma2))))`

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE 2

Nella rimanente parte di questa sezione si trasforma l'equazione dell'ellisse nella forma canonica tradizionale. Utilizzando la scomposizione spettrale, e le proprietà del prodotto di matrici inverse ed il fatto che la matrice V è ortogonale, l'equazione (3.11) può essere riscritta come

$$\begin{aligned}
 (X_1 - \bar{x}_1 \quad X_2 - \bar{x}_2) (V \Lambda V')^{-1} \begin{pmatrix} X_1 - \bar{x}_1 \\ X_2 - \bar{x}_2 \end{pmatrix} &= 1 \\
 (\tilde{X}_1 \quad \tilde{X}_2) (V \Lambda V')^{-1} \begin{pmatrix} \tilde{X}_1 \\ \tilde{X}_2 \end{pmatrix} &= 1 \\
 (\tilde{X}_1 \quad \tilde{X}_2) (V \Lambda^{-1} V') \begin{pmatrix} \tilde{X}_1 \\ \tilde{X}_2 \end{pmatrix} &= 1 \\
 (Y_1 \quad Y_2) \Lambda^{-1} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} &= 1 \\
 \frac{Y_1^2}{\lambda_1} + \frac{Y_2^2}{\lambda_2} &= 1 \tag{3.12}
 \end{aligned}$$

dove

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = V' \begin{pmatrix} X_1 - \bar{x}_1 \\ X_2 - \bar{x}_2 \end{pmatrix} = V' \begin{pmatrix} \tilde{X}_1 \\ \tilde{X}_2 \end{pmatrix}$$

L'espressione (3.12) non è altro che l'equazione dell'ellisse nel sistema di assi (Y_1, Y_2) in forma canonica⁵. La trasformazione dalle coordinate X_1, X_2 , alle coordinate \tilde{X}_1, \tilde{X}_2 non è altro che una traslazione degli assi cartesiani. La trasformazione dalle coordinate \tilde{X}_1, \tilde{X}_2 (in funzione della base canonica e_1, e_2) alle coordinate Y_1, Y_2 in funzione della nuova base determinata da v_1 e v_2

⁵L'equazione canonica dell'ellisse si riferisce ad una ellisse che ha centro nell'origine degli assi e semiassi posti lungo gli assi cartesiani.

(nuovi assi cartesiani) non è altro che una rotazione degli assi. La Figura 3.6 contiene la rappresentazione dei punti originali in termini di scostamenti dalla media. In questa figura i punti fuori dall'ellisse sono stati etichettati. Il codice per ottenere questa figura è riportato di seguito:

```
%% Rappresentazione dei punti in termini di scostamenti dalla media
% Punti nella base canonica e_1 e e_2
figure
Xtilde1=Xtilde(:,1);
Xtilde2=Xtilde(:,2);

plot(Xtilde1,Xtilde2,'o')
xlabel('$\tilde{X}_1=\overline{x}_1$', 'Interpreter', 'latex')
ylabel('$\tilde{X}_2=\overline{x}_2$', 'Interpreter', 'latex')
hold('on')
Ell= ellipse(0,S,1);
axis equal
[in]=inpolygon(Xtilde1,Xtilde2,Ell(:,1),Ell(:,2));
seq=1:n;
textOut=string(seq(~in));
text(Xtilde1(~in),Xtilde2(~in),textOut)
```

La Figura 3.7 contiene la rappresentazione dei punti dopo la rotazione nel nuovo sistema di assi cartesiani determinato da v_1 e v_2 . Anche in questa figura i punti fuori dall'ellisse sono stati etichettati. Si noti che in questa trasfor-

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE 2

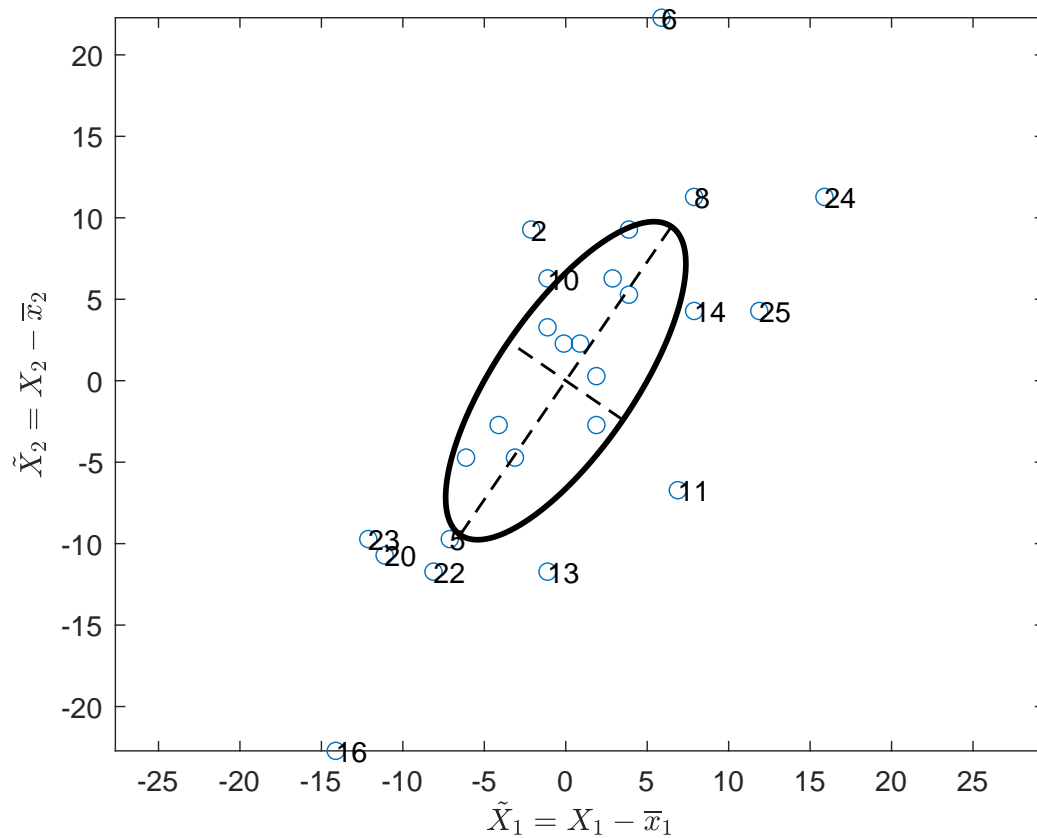


Figura 3.6: Diagramma di dispersione in termini di scostamenti dalla media. L'ellisse è il luogo dei punti $(\tilde{X}_1 \ \tilde{X}_2)S^{-1}(\tilde{X}_1 \ \tilde{X}_2)' = 1$

mazione sia la traccia sia il determinante della forma quadratica rimangono invariati: $tr(S) = tr(\Lambda) = \sum_{i=1}^p \lambda_i$ e $det(S) = det(\Lambda) = \prod_{i=1}^p \lambda_i$.

Il codice per ottenere la Figura 3.7 è riportato di seguito

```
figure
Y=Xtilde*V;
plot(Y(:,1),Y(:,2),'o')
hold('on')
Ellr= ellipse(0,Lambda,1);
xlabel('$Y_1$=Prima componente principale','Interpreter','latex')
```

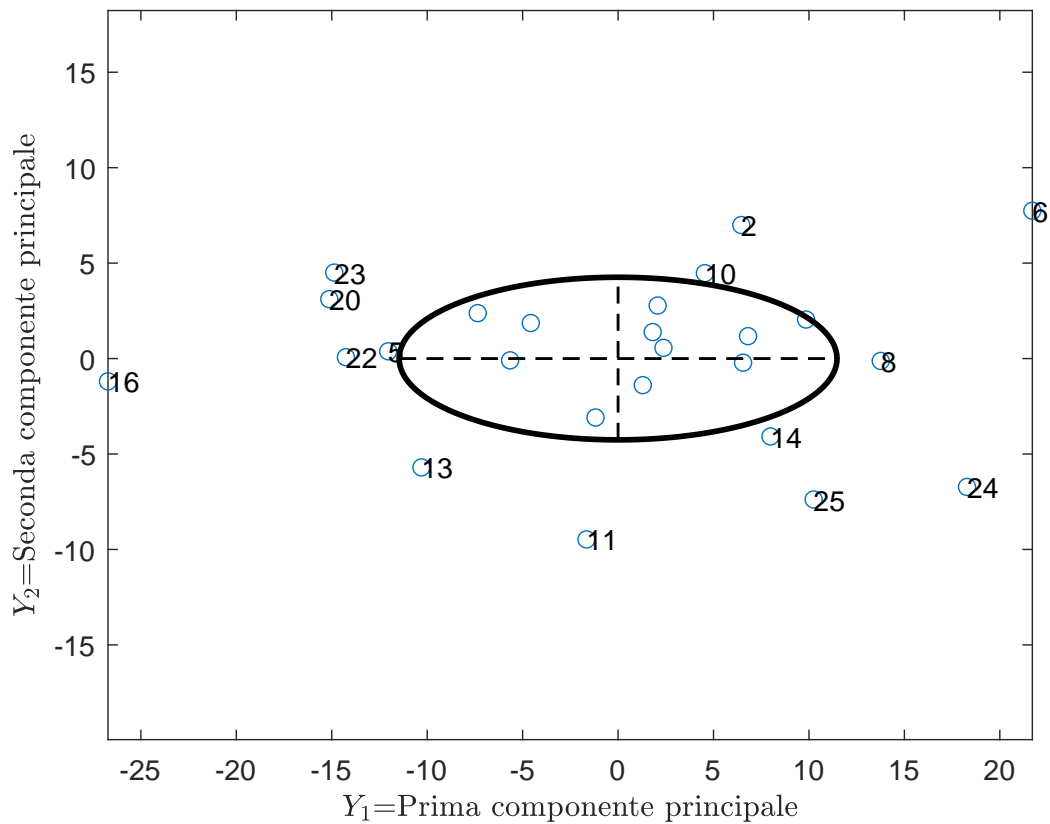


Figura 3.7: Diagramma di dispersione in termini delle prime due componenti principali Y_1 e Y_2 . L'ellisse è il luogo dei punti $\frac{Y_1^2}{\lambda_1} + \frac{Y_2^2}{\lambda_2} = \frac{Y_1^2}{131.5183} + \frac{Y_2^2}{18.135} = 1$

```
ylabel('$Y_2$=Seconda componente principale','Interpreter','latex')
```

```
[in]=inpolygon(Y(:,1),Y(:,2),Ellr(:,1),Ellr(:,2));
seq=(1:size(X,1))';
textOut=string(seq(~in));
text(Y(~in,1),Y(~in,2),textOut)
axis equal
```

In conclusione, la trasformazione in componenti principali $Y = \tilde{X}V$ (oppure ZV) equivale ad una semplice rotazione degli assi cartesiani. Nello spa-

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

zio Y_1, \dots, Y_p le variabili Y hanno matrice di covarianze $cov(Y) = \Lambda$. Nello spazio $\tilde{X}V\Lambda^{-1/2} = \tilde{X}V\Gamma^{-1}$ le componenti principali sono standardizzate e $cov(\tilde{X}V\Gamma^{-1}) = I$ (v. equazione 3.9). Le coordinate dell'ellisse diventano un cerchio (v. Figura 3.8).

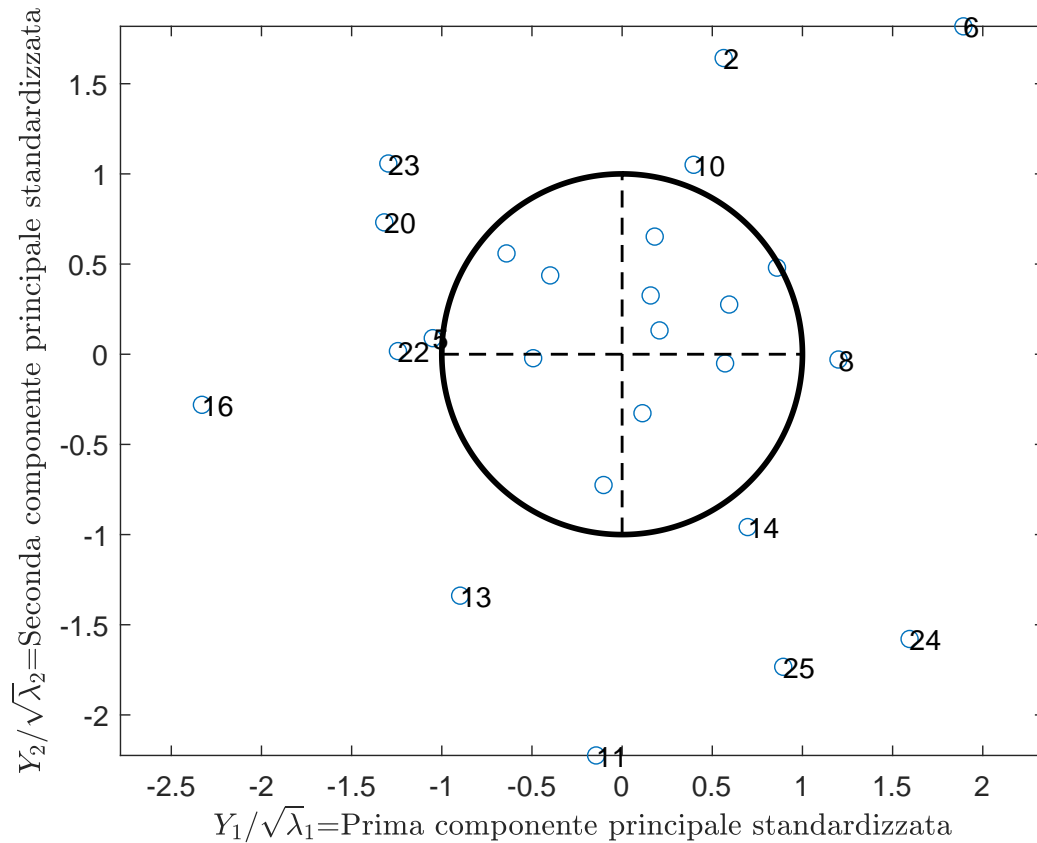


Figura 3.8: Diagramma di dispersione in termini delle prime due componenti principali standardizzate $Yst_1 = Y_1/\sqrt{\lambda_1}$ e $Yst_2 = Y_2/\sqrt{\lambda_2}$. La circonferenza è il luogo dei punti $Yst_1^2 + Yst_2^2 = 1$

Il codice per ottenere la Figura 3.8 è riportato di seguito

```
figure
```

```
Yst=zscore(Y);
```

```
plot(Yst(:,1),Yst(:,2),'o')
```

```

hold('on')
Ellr= ellipse(0,eye(2),1);
xlabel(['$Y_1/\sqrt{\lambda_1}$=Prima componente principale ' ...
       'standardizzata'],'Interpreter','latex')
ylabel(['$Y_2/\sqrt{\lambda_2}$=Seconda componente principale' ...
       ' standardizzata'],'Interpreter','latex')

[in]=inpolygon(Yst(:,1),Yst(:,2),Ellr(:,1),Ellr(:,2));
textOut=string(seq(~in));
text(Yst(~in,1),Yst(~in,2),textOut)
axis equal

```

Il riassunto di questa sezione e di quella precedente è il seguente: il vettore $\tilde{X}v_j$, $j = 1, 2, \dots, p$ contiene le coordinate dei punti $\tilde{x}_1, \dots, \tilde{x}_n$ proiettati nello spazio della j -esima componente principale. $Xv_jv'_j$ contiene invece i punti rappresentati nello spazio di partenza (dove i vettori sono rappresentati nella base canonica I_p). La scrittura

$$\tilde{X}VV' = YV'$$

dove V è una matrice ortogonale, significa che il nuovo sistema di assi cartesiani è determinato dalle colonne di V (nuova base ortogonale). Y contiene le coordinate dei punti ruotati nel nuovo sistema di assi cartesiani di V . In generale quindi, la scrittura

$$\tilde{X} = GH' \tag{3.13}$$

3.6. PC COME PROIEZIONE ORTOGONALE DEI PUNTI IN UN SOTTOSPAZIO DI DIMENSIONE

dove H è una matrice ortogonale, significa che i punti $\tilde{x}_1, \dots, \tilde{x}_n$ sono rappresentati nella base determinata dalle colonne di H . Le coordinate di questi punti nella base H sono date da G . Quindi, nella svd della matrice \tilde{X} l'espressione:

$$\tilde{X} = U\Gamma V' = GV'$$

indica che $G = U\Gamma = \tilde{X}V$ contiene le coordinate delle righe di \tilde{X} rispetto alla base V (v. anche equazione (1.12)). Ogni riga di \tilde{X} può essere espressa come (v. anche equazione (1.11)):

$$\tilde{x}_i = \sum_{j=1}^p g_{ij}v_j$$

Di conseguenza, l' i -esima riga di G contiene le coordinate di \tilde{x}_i nella base V . Similmente, la matrice $H = V\Gamma$ contiene le coordinate delle colonne di \tilde{X} nello spazio determinato dalle colonne della matrice U .

Nella scomposizione

$$\tilde{X} = U\Gamma V' = UH'$$

se facciamo la trasposta

$$\tilde{X}' = H_{p \times p}(U_{n \times p})'$$

ogni colonna \tilde{X}_r di \tilde{X} può essere espressa come combinazione lineare dei vettori di base u_1, \dots, u_p .

$$\tilde{X}_j = \sum_{i=1}^p h_{ij}u_i$$

3.7 L'analisi in componenti principali in pratica

L'obiettivo di questa sezione è quello di mostrare attraverso un dataset reale le varie tappe dell'analisi in componenti principali. A titolo di esempio consideriamo i dati della qualità della vita di 107 province italiane riportati nel file `benessere.xlsx`. In questo caso, dato che le variabili originarie presentano diversa unità di misura e diverso ordine di grandezza è necessario passare alla matrice degli scostamenti standardizzati Z .

Il requisito di base per applicare l'analisi in componenti principali sta nel fatto che le variabili analizzate siano correlate tra loro. Le variabili che presentano elevati valori della correlazione in modulo possono esprimere, almeno in parte, lo stesso tipo di informazione, che può essere isolata e rappresentata tramite la costruzione di variabili di sintesi. È necessario, quindi, costruire la matrice di correlazione ed esaminare l'entità delle stesse. Il codice per caricare i dati e costruire la matrice di correlazione è riportato di seguito

```
%% CARICAMENTO DATI

Xtable=readtable('benessere.xlsx','ReadRowNames',true);

% X = matrice di double senza nomi delle righe e nomi delle colonne
X=table2array(Xtable);

% nameXvars = cell che contiene i nomi delle variabili
nameXvars=Xtable.Properties.VariableNames;
```

```

[n,p]=size(X);

%% Calcolare la matrice di correlazione,
% Standardizzo i dati
Z=zscore(X);
% calcolo matrice di correlazione
R=cov(Z);
% mostro la matrice di correlazione in formato table
Rtable=array2table(R,'RowNames',nameXvars,'VariableNames',nameXvars);
format bank
disp(Rtable)
format short
% R poteva essere ottenuta direttamente da X utilizzando la funzione corr

```

Il codice sopra produce

```

      va    depos pensioni disocc export fallimen protesti
      ----  -
va          1.00  0.89   0.74  -0.80   0.56   0.06  -0.37
depos       0.89  1.00   0.73  -0.69   0.44   0.17  -0.24
pensioni    0.74  0.73   1.00  -0.50   0.31   0.35  -0.29
disocc      -0.80 -0.69  -0.50   1.00  -0.61   0.15   0.40
export       0.56  0.44   0.31  -0.61   1.00   0.04  -0.22
fallimen     0.06  0.17   0.35   0.15   0.04   1.00   0.46
protesti    -0.37 -0.24  -0.29   0.40  -0.22   0.46   1.00

```

Le correlazioni tra alcune coppie di variabili (ad esempio tra *va* e depositi, oppure tra *va* e pensioni) sono particolarmente elevate, di conseguenza è lecito ipotizzare che questi dati possano essere rappresentati in uno spazio a ridotta dimensione con una perdita limitata di informazione.

Il passo successivo consiste nel calcolare gli autovalori ed autovettori della matrice R ed ottenere la matrice delle componenti principali tramite la trasformazione $Y = ZV$. Il codice per effettuare queste operazioni è riportato di seguito.

```
%% Autovettori e autovalori della matrice di correlazione
[V,La]=eig(R);
la=diag(La);
[aa,indsor]=sort(la,'descend');
% Riordino le colonne della matrice V e La in modo tale che
% La(1,1) sia il grande autovalore e V(:,1) sia l'autovettore associato
% La(2,2) sia il secondo più grande autovalore
% e V(:,2) sia l'autovettore associato
V=V(:,indsor);
lasor=la(indsor);
La=diag(lasor);

% CALCOLO COMPONENTI PRINCIPALI
Y=Z*V;
```

A questo punto è necessario analizzare la quota di varianza relativa e cumulata spiegata dalla diverse componenti principali (ossia dalle diverse

colonne della matrice Y). Occorre ricordare che le variabili sono state standardizzate, di conseguenza la varianza totale è pari p . Tenendo presente l'equazione (3.1), la tabella che riporta il quadro della varianza spiegata dalle diverse dimensioni latenti può essere ottenuta tramite il codice che segue:

```
% namePCS contiene la sequenza di etichette PC1, ..., PCp
namePCS=cellstr([repmat('PC',p,1) num2str((1:p)')]);
% Calcolo la quota di varianza spiegata da ciascuna CP
% autoval è una matrice di dimensione px3 che contiene
% Prima colonna: autovalori ordinati
% Seconda colonna: 100*autovalori/p
% Terza colonna: somma cumulata in percentuale
autoval=[lasor 100*(lasor)/p 100*cumsum(lasor)/p];
namecols={'Autovalori' 'Var_spiegata' 'Var_cum_spiegata'};
autovaltable=array2table(autoval,'RowNames', ...
    namePCS,'VariableNames',namecols);
disp(autovaltable)
```

L'ultima istruzione `disp(autovaltable)` produce

	Autovalori	Var_spiegata	Var_cum_spiegata
	-----	-----	-----
PC1	3.735	53.358	53.358
PC2	1.5404	22.006	75.364
PC3	0.77724	11.103	86.467
PC4	0.48173	6.8818	93.349

PC5	0.20482	2.926	96.275
PC6	0.18137	2.591	98.866
PC7	0.079388	1.1341	100

Le percentuali singole e cumulate spiegate dalle diverse componenti può anche visualizzata tramite diagramma di Pareto come segue:

```
figure
pareto(autoval(:,1),namePCs)
xlabel('Componenti principali')
ylabel('Varianza spiegata (%)')
```

L'output di questo codice è riportato nella Figura 3.9. La sostituzione alle 7 variabili originarie delle prime due componenti principali, comporta una perdita di informazione di poco inferiore al 25 per cento. Esistono 4 criteri da seguire che possono aiutare a capire quante componenti principali si devono scegliere.

- Varianza spiegata. L'analisi per essere valida deve spiegare almeno il 60-70 per cento di varianza delle variabili originarie.
- Spiegazione in media del 95 per cento di varianza di ogni variabile e complessivamente una quota pari a 0.95^p . In questo esempio questo criterio afferma che è necessario considerare un numero di componenti che spieghino ($0.95^7 = 0.6983$) almeno il 69.83 per cento di variabilità delle variabili originarie.
- Criterio dell'eigenvalue maggiore di uno (regola di Kaiser). Dato che le variabili sono standardizzate è necessario considerare gli autovalo-

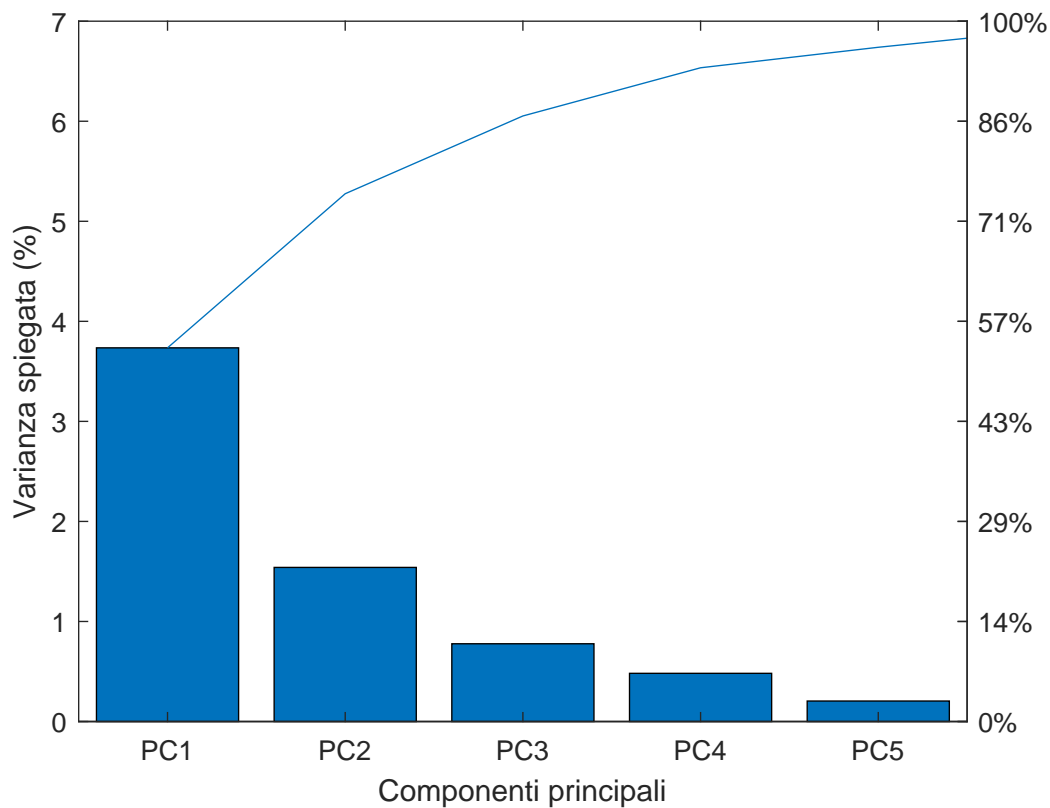


Figura 3.9: Diagramma di Pareto della percentuale di varianza spiegata dalle diverse componenti principali.

ri maggiori di uno, poiché in questo caso ogni componente principale spiega di più di una singola variabile.

- Criterio dello scree plot (o diagramma a falda) ossia il grafico che unisce tramite linea la percentuale spiegata dalle diverse componenti. L'idea è quella di scegliere il numero di componenti corrispondente al gomito della spezzata.

Nel nostro caso, tutti i criteri di cui sopra portano a considerare solo per le prime due componenti principali.

Il passo successivo consiste nell'interpretare le componenti principali che si è deciso di estrarre. A questo proposito è necessario costruire la matrice di correlazione tra le variabili originarie e le componenti principali che sono state estratte. La matrice di covarianza tra X e Y può essere scritta come segue

$$\text{cov}(X, Y) = \text{cov}(\tilde{X}, Y) = \tilde{X}'Y/(n-1) = \tilde{X}'\tilde{X}V/(n-1) = SV = V\Lambda.$$

Il coefficiente di correlazione tra le p variabili originarie e la j -esima componente principale è dato da;

$$r_{X,y_j} = \frac{v_j \lambda_j D^{-1}}{\sqrt{\lambda_j}} = v_j \sqrt{\lambda_j} D^{-1}$$

dove D è la matrice diagonale che contiene sulla diagonale principale gli scostamenti quadratici medi delle variabili originali (v. sezione 1.8).

Se le variabili originarie sono state standardizzate $D_j = I_p$ e otteniamo che

$$r_{X,y_j} = v_j \sqrt{\lambda_j}$$

La matrice che contiene le correlazioni tra le variabili e le componenti principali è solitamente denominata matrice di componenti o matrice dei loadings. Il codice per ottenere questa matrice con riferimento alle prime due PC è riportato di seguito

```
MatrComp=V*sqrt(La);
% MatrComp ha dimensione $p \times p$. Nel nostro esempio 7*7;
% Ora la ridefinisco prendendo solo le correlazioni tra le variabili
```

```
% originarie e le prime due componenti principali
MatrComp=MatrComp(:,1:2);

%% Matrice di componenti in formato table
MatrCompt=array2table(MatrComp,"RowNames",nameXvars,"VariableNames",namePCs(1:2));
disp('Correlazioni tra le variabili originarie e le PC')
disp(MatrCompt)
```

Questo codice produce

	PC1	PC2
	-----	-----
va	-0.95817	0.041531
depos	-0.89254	0.19503
pensioni	-0.78774	0.33857
disocc	0.86247	0.21298
export	-0.66304	-0.059315
fallimen	-0.062234	0.93191
protesti	0.46092	0.6846

La prima componente principale è correlata in maniera forte ed inversa con le variabili va, deposit, pensioni ed export ed è correlata positivamente con il tasso di disoccupazione (disocc) ed in misura minore con i protesti. Questa variabile latente può essere interpretate come un indicatore complessiva del grado di povertà della provincia. In altri termini, tanto più grande sarà il

punteggio sulla prima componente principale (Y_1) tanto più povera sarà la provincia.

La seconda componente principale è correlata in maniera forte e diretta con le variabili fallimenti e protesti. Questa componente, quindi, deve essere interpretata come un indicatore di “malessere delle aziende”. Tanto più è alto il valore di Y_2 tanto più elevato sarà l’indice del malessere delle aziende per quella determinata provincia.

Osservazione: gli autovettori sono definiti a meno del segno (v. equazione 1.16) di conseguenza se si scambiano di segno gli elementi di v_1 la prima componente principale deve essere interpretata come indice di ricchezza.

Tenendo presente che la colonna j della matrice di componenti è $v_j\sqrt{\lambda_j}$, è immediato verificare che la somma dei quadrati di ogni colonna è pari al rispettivo autovalore λ_j , $j = 1, 2, \dots, p$:

$$(v_j\sqrt{\lambda_j})'(v_j\sqrt{\lambda_j}) = \lambda_j v_j' v_j = \lambda_j$$

Il codice per verificare questa uguaglianza è riportato di seguito:

```
% La somma dei quadrati della colonna j della matrice di componenti è
% pari al j-esimo autovalore
j=1;
disp(['La somma dei quadrati della colonna ' num2str(j) ' della matrice di co
sum(MatComp(:,j).^2)
disp(['è uguale all'' autovalore ' num2str(j) '=' num2str(lasor(j))])
```

Le correlazioni tra le variabili originarie e le componenti principali possono essere visualizzate tramite grafici a barre (v. Figura 3.10).

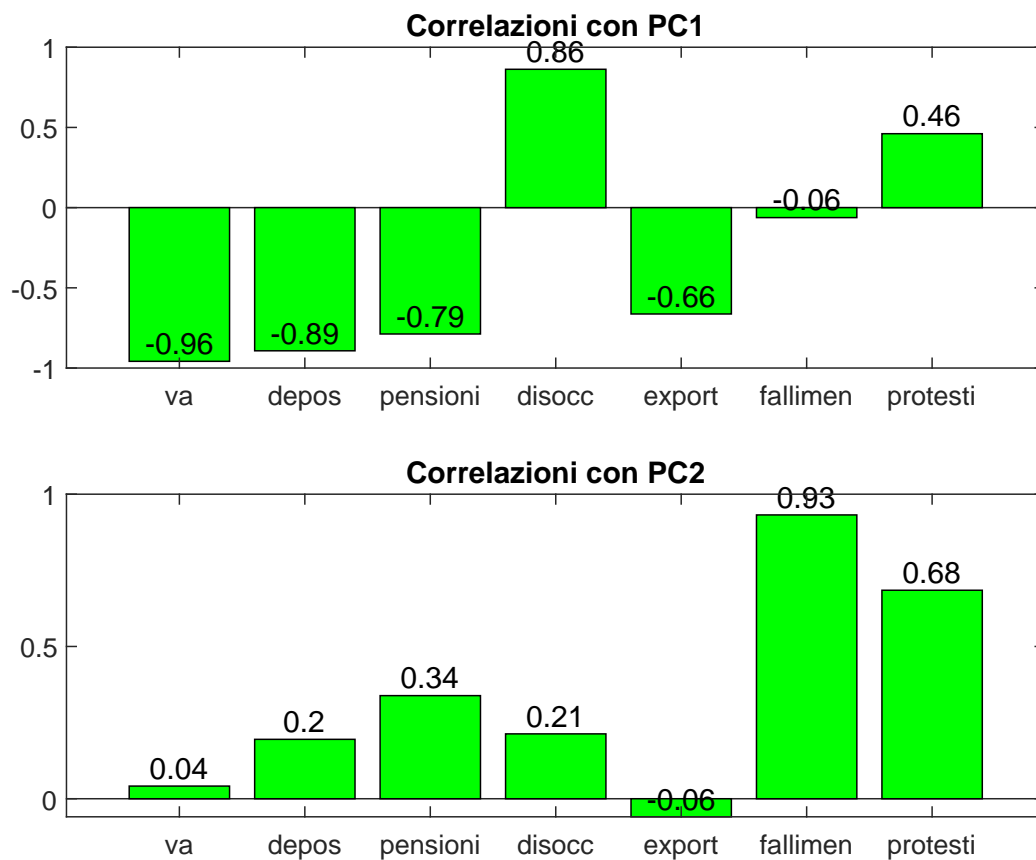


Figura 3.10: Correlazioni tra le variabili originarie e la prima componente principale (pannello in alto) e la seconda componente principale (pannello in basso).

La somma dei quadrati di ogni riga della matrice di componenti è la cosiddetta *comunalità* (*communality*) ossia la quota di varianza di ogni variabile spiegata dalle prime k componenti (se abbiamo deciso di avere k componenti). Il codice per ottenere le comunalità è riportato di seguito:

```
disp(['Comunalità: quote di varianza di ogni ' ...
      'variabile spiegate dalle prime due CP'])
Comu=sum(MatComp.^2,2);
```

```
Comutable=array2table(Comu,'RowNames',nameXvars,...
    'VariableNames',{'Comunalità'});
disp(Comutable)
```

Questo codice produce il seguente output:

Comunalità: quote di varianza di ogni variabile spiegate dalle prime due CP

```
Comunalità
-----
va          0.91982
depos       0.83466
pensioni    0.73516
disocc      0.78921
export      0.44314
fallimen    0.87234
protesti    0.68113
```

La variabile che viene spiegata meglio è va, seguita da fallimen. Quella spiegata peggio è export.

Il passo successivo consiste nel rappresentare tramite frecce i numeri che si trovano nella matrice di componenti. Il codice che segue

```
zeroes = zeros(p,1);
% Frecce che partono dall'origine e arrivano fino a MatrComp
quiver(zeroes,zeroes,MatrComp(:,1),MatrComp(:,2))
% Label delle frecce
```

```

text(MatComp(:,1),MatComp(:,2),nameXvars,...
      'VerticalAlignment','bottom','HorizontalAlignment','center');
% Vengono aggiunti gli assi cartesiani
xline(0)
yline(0)
% Aggiunta delle label sugli assi
xlabel('Prima PC: Indice di povertà');
ylabel('Seconda PC: Indice di malessere delle aziende');
% Stessa unità di misura per i due assi
axis equal

```

produce come output la Figura 3.11. Tenendo presente i concetti visti nella sezione 1.1.6 (prodotto scalare tra due vettori) e nella sezione 1.6 (proiezioni ortogonali) emerge immediatamente che:

- Il coseno dell'angolo tra due vettori indica la correlazione tra le due variabili corrispondenti. Variabili altamente correlate puntano nella stessa direzione (v. ad esempio *va* e *depos* oppure *va* e *export*), variabili poco correlate sono ad angolo retto (v. ad esempio *fallim* e *disocc*).
- Se il coseno dell'angolo tra due vettori è vicino a 180 gradi significa che le due variabili presentano una forte correlazione negativa (v. ad esempio *va* e *disocc* oppure *export* e *disocc*).
- L'angolo vicino a zero tra un vettore ed un asse cartesiano significa che la variabile è fortemente correlata con quella dimensione latente (v. ad esempio *fallimen* con la seconda componente principale). L'angolo

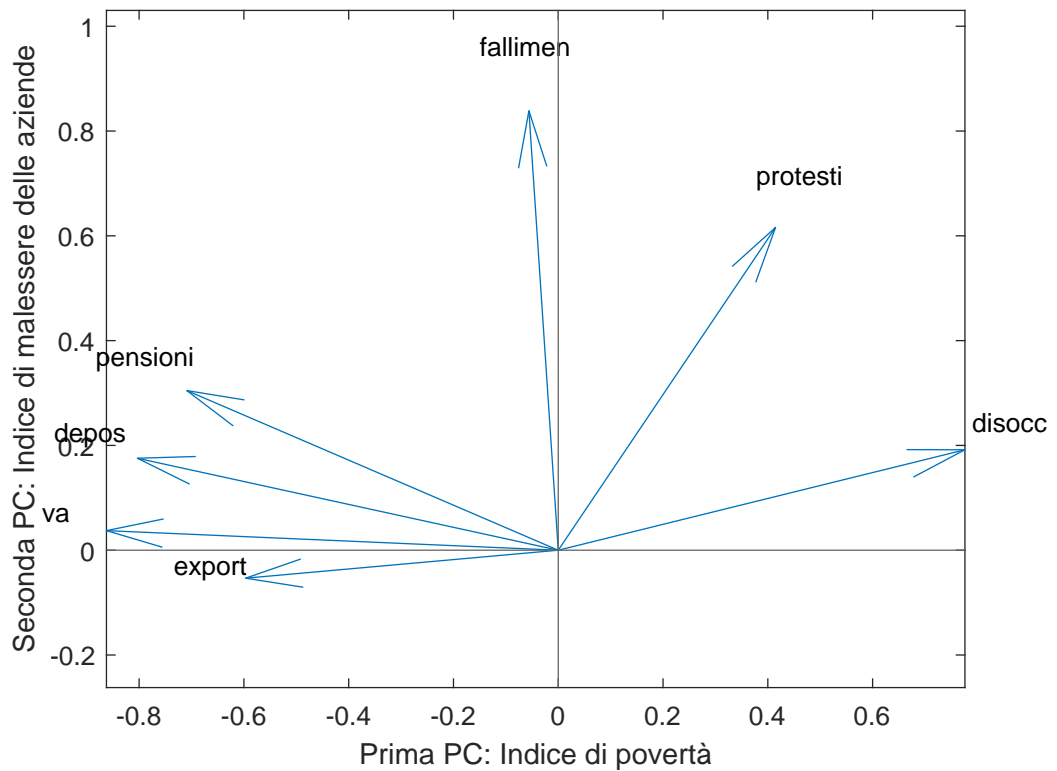


Figura 3.11: Rappresentazione grafica delle correlazioni tra le variabili originarie e le prime due componenti principali tramite vettori (freccie)

vicino a 90 gradi tra un vettore ed un asse significa che la variabile non è correlata con quell'asse (ad esempio *va* con la seconda componente principale).

- il coseno dell'angolo tra un vettore e un asse principale è proporzionale al contributo della variabile corrispondente alla determinazione della variabilità totale dell'asse espressa dall'autovalore. Più precisamente dato che $\sum_{i=1}^p (v_{ij}\sqrt{\lambda_j})^2 = \lambda_j$, dove v_{ij} è l'elemento i -esimo dell'autovettore j -esimo, il contributo della variabile i alla variabilità dell'asse j -esimo è data

$$v_{ij}^2 \lambda_j / \lambda_j = v_{ij}^2$$

il quadrato dell'elemento i -esimo dell'autovettore j -esimo. Ad esempio, dall'esame della Figura 3.11 emerge che va gioca un ruolo importante nella determinazione della prima componente principale.

- la lunghezza (norma) di ogni vettore corrisponde alla radice quadrata della comunalità. Tanto più la lunghezza della freccia si avvicina ad 1, tanto meglio la variabile è spiegata dalle prime due componenti principali. Dato che la freccia con la lunghezza più piccola è associata alla variabile `export` significa che questa variabile è quella con la comunalità più bassa.

Il passo successivo consiste nell'analizzare e rappresentare graficamente gli scores Y_1 Y_2 delle prime due componenti principali. Il codice che segue

```
plot(Y(:,1),Y(:,2),'o')
xlabel('Prima PC: Indice di povertà');
ylabel('Seconda PC: Indice di malessere delle aziende');
text(Y(:,1),Y(:,2),Xtable.Properties.RowNames)
% Vengono aggiunti gli assi cartesiani
xline(0)
yline(0)
```

produce il grafico mostrato nella Figura 3.12. Procedendo da sinistra verso destra l'indice di povertà aumenta. Similmente procedendo dal basso verso l'alto l'indice di malessere delle aziende aumenta. Le province dove si vive meglio, quindi, sono quelle che si collocano nel terzo quadrante. Punti vicini

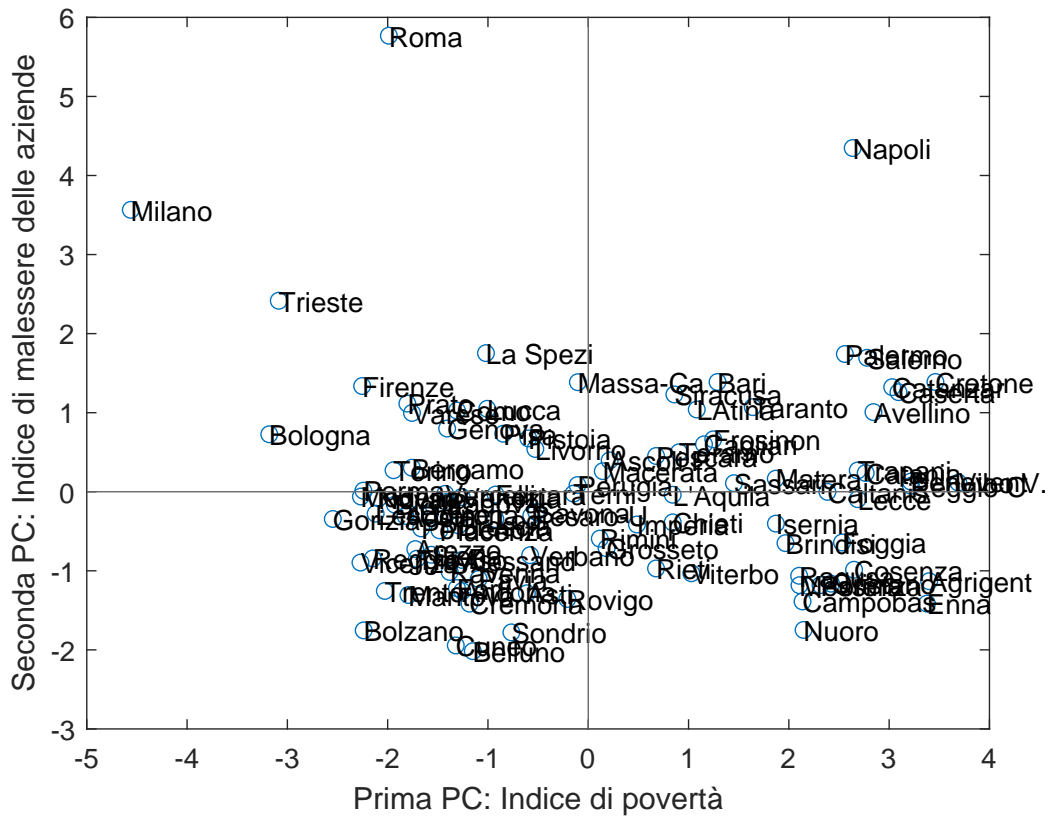


Figura 3.12: Rappresentazione grafica degli scores non standardizzati delle prime due componenti principali tramite punti

rappresentano province con caratteristiche simili. In questa rappresentazione $var(Y_1) = \lambda_1$ e $var(Y_2) = \lambda_2$. Dato che $\sum_{i=1}^n y_{ij}^2 / (n-1) = \lambda_j$,

$$\frac{y_{ij}^2}{(n-1)\lambda_j}$$

rappresenta il contributo dato dalla i -esima coordinata $i = 1, 2, \dots, n$ alla determinazione della variabilità dell'asse j -esimo. Dunque, tanto più è alto il quadrato della coordinata, tanto più è alto il contributo alla variabilità dell'asse. Ad esempio, dall'esame della Figura 3.12 emerge che Milano (città più ricca) è il punto che contribuisce di più alla determinazione del primo

autovalore (ossia alla variabilità della prima dimensione latente associata all'indice di povertà). Roma e Napoli (le due città con il più alto valore di malessere delle aziende), al contrario, sono i due punti che contribuiscono di più alla determinazione del secondo autovalore (ossia alla variabilità della seconda dimensione latente associata all'indice di malessere delle aziende). Il coseno al quadrato tra i singoli punti riga ed un determinato asse è una misura della qualità della rappresentazione dei punti unità nel sottospazio generato dai fattori latenti, in quanto tanto più il coseno al quadrato risulta vicino ad 1, tanto più il punto in proiezione avrà conservato la distanza dall'origine. In altri termini il coseno al quadrato (v. Figura 1.10) è il rapporto di due lunghezze al quadrato, il cateto (che rappresenta la lunghezza della proiezione sull'asse) e l'ipotenusa (che rappresenta la posizione originale del punto nello spazio a p dimensioni). Un valore del coseno vicino a zero significa che la lunghezza del cateto è vicina a quella dell'ipotenusa e che quindi il punto è ben rappresentato dalla corrispondente dimensione latente.

La rappresentazione degli scores standardizzati con varianza unitaria $Y_1/\sqrt{\lambda_1}$ e $Y_2/\sqrt{\lambda_2}$ è riportata nella figura 3.13. Il codice che ha generato questa figura è riportato di seguito.

```
% Yst= componenti principali standardizzate cov(Yst) = matrice identità
Yst=Y*sqrt(inv(La));

% Rappresentazione nel piano cartesiano degli scores (standardizzati)
plot(Yst(:,1),Yst(:,2),'o')

xlabel('Prima PC: Indice di povertà');

ylabel('Seconda PC: Indice di malessere delle aziende');

text(Y(:,1),Y(:,2),Xtable.Properties.RowNames)
```

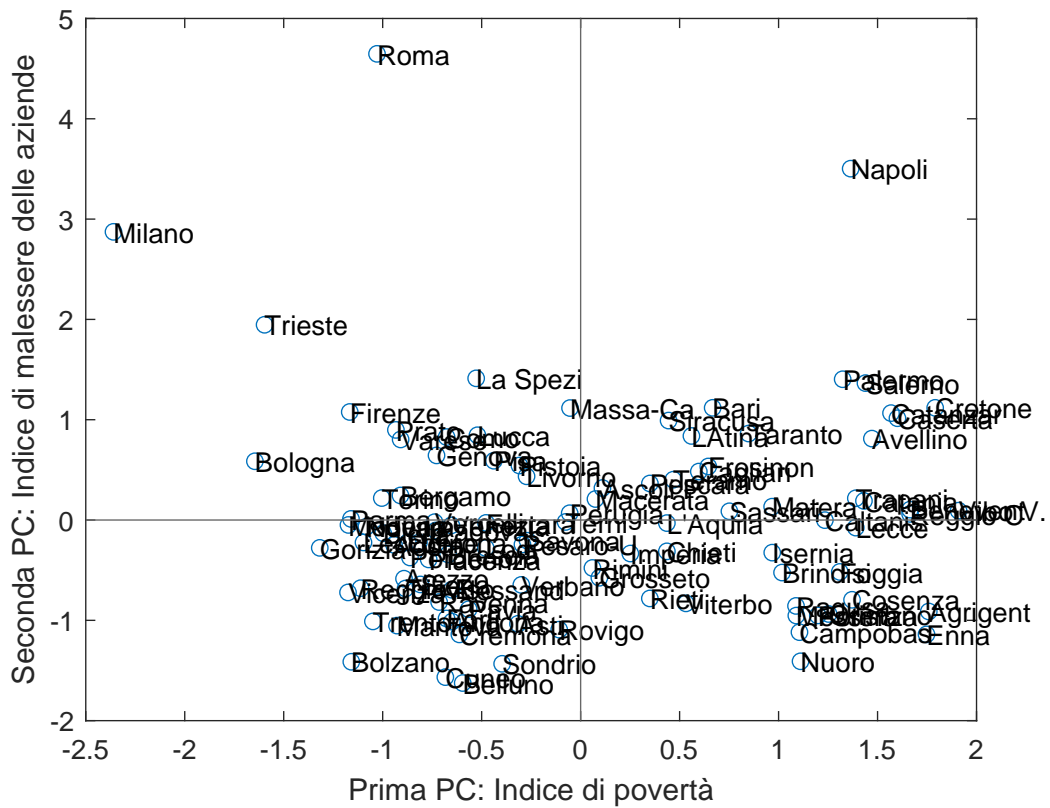


Figura 3.13: Rappresentazione grafica degli scores standardizzati delle prime due componenti principali tramite punti

% Vengono aggiunti gli assi cartesiani

xline(0)

yline(0)

3.8 Il biplot

Il biplot è una rappresentazione simultanea delle n unità (righe della matrice dei dati) e delle p variabili (colonne della matrice dei dati) in uno spazio a due dimensioni.

L'idea sottostante al biplot consiste nell'aggiungere l'informazione sulla

relazione tra variabili al grafico degli scores delle componenti principali. Il suffisso bi indica le due informazioni contenute in X e rappresentate nel grafico: le righe di X rappresentano le osservazioni campionarie, le colonne di X rappresentano le variabili. La costruzione del biplot si basa sulla scomposizione in valori singolari della matrice dei dati in termini di scostamenti dalla media \tilde{X} o della matrice degli scostamenti standardizzati Z . In questa sezione facciamo riferimento alla matrice Z (di dimensione $n \times p$) degli scostamenti standardizzati. Ovviamente è possibile partire anche dalla matrice \tilde{X} degli scostamenti dalla media. In tal caso i richiami alla matrice di correlazione $R = Z'Z/(n-1)$ devono intendersi riferiti alla matrice di covarianze $S = \tilde{X}'\tilde{X}/(n-1)$.

Si parte dalla scomposizione in valori singolari che abbiamo introdotto nella sezione 3.4:

$$Z = U\Gamma V'$$

Se le prime due componenti principali tengono conto di una quota elevata della varianza totale è possibile sostituire alla matrice Z la sua miglior rappresentazione di rango 2 come segue:

$$Z \approx U_{(2)}\Gamma_{(2)}^*V_{(2)}'$$

dove $U_{(2)}$ è una matrice di dimensione $n \times 2$ che contiene le prime due colonne della matrice U , $V_{(2)}$ è una matrice di dimensione $p \times 2$ che contiene le prime due colonne della matrice V . $\Gamma_{(2)}^*$ è la matrice diagonale di dimensione 2×2 che contiene sulla diagonale principale i primi due valori singolari della matrice Z (ossia la radice quadrata degli autovalori della matrice $Z'Z =$

$(n-1)R^6$. Per fare in modo che questa matrice contenga gli autovalori di $Z'Z/(n-1) = R$, possiamo scrivere la scomposizione in valori singolari come segue:

$$\frac{Z}{\sqrt{n-1}} \approx U_{(2)} \frac{\Gamma_{(2)}^*}{\sqrt{n-1}} V_{(2)}'$$

$$Z \approx \hat{Z} = \sqrt{n-1} U_{(2)} \frac{\Gamma_{(2)}^*}{\sqrt{n-1}} V_{(2)}'$$

$$Z \approx \hat{Z} = \sqrt{n-1} U_{(2)} \Gamma_{(2)} V_{(2)}'$$

dove $\Gamma_{(2)} = \Gamma_{(2)}^* / \sqrt{n-1}$.

Per illustrare i concetti di cui sopra facciamo riferimento al dataset delle lavatrici riportato nella Tabella 3.2.

Il codice per ottenere le matrici descritte sopra è riportato di seguito.

```
Xtable=readtable('lavatrici.xlsx','ReadRowNames',true,'Sheet','dati');
% X = matrice di double senza nomi delle righe e nomi delle colonne
X=table2array(Xtable);
[n,p]=size(X);

% Standardizzo i dati
Z=zscore(X);

% svd
[U,Gammastar,V]=svd(Z,'econ');

% Controllo sulla svd
```

⁶Se si desidera visualizzare graficamente le dimensioni di queste matrici si può fare riferimento alla Figura 3.17.

marca	prezzo	giri	cons	profon	energia	load
LG ELECTRONICS	799	1400	75	60	133	70
WHIRLPOOL	671	1200	49	60	114	60
MIELE	1469	1600	44	58	95	50
INDESIT	329	600	64	53	95	50
PHILCO	450	1000	65	60	109	60
SAMSUNG	369	800	49	55	120	50
IGNIS	309	600	64	51	95	50
AEG	950	1600	42	60	94	50
ELECTROLUX	759	1600	44	58	95	50
SIEMENS	798	1000	39	55	95	60
ARISTON	488	1000	59	54	95	55
REX	649	1200	44	60	95	50
BOSCH	671	1000	49	55	95	50
CANDY	449	700	58	54	95	50
SAN GIORGIO	449	1100	59	52	114	50
ZOPPAS	389	650	64	54	95	50

Tabella 3.2: Matrice dei dati riferita a 16 lavatrici. Sono state rilevate 6 variabili

```

maxdiff=max(abs(Z-U*Gammastar*V'), [], "all");

assert(maxdiff<1e-12,"Errore di programmazione. " + ...

    "La matrice originaria non è stata ricostruita")

sqn1=sqrt(n-1);

Gamma=Gammastar/sqn1;

% Ricostruzione della matrice originaria

% con una matrice di rango numcomp

numcomp=2;

U2=U(:,1:numcomp);

V2=V(:,1:numcomp);

Gamma2=Gamma(1:2,1:2);

Zhat=sqn1*U2*Gamma2*V2';

```

La matrice Z degli scostamenti standardizzati risulta, quindi, approssimativamente uguale al prodotto di due matrici, la prima (G) riferita alle n unità (di dimensione $n \times 2$) e la seconda (H) alle p variabili (di dimensione $p \times 2$) come segue:

$$Z_{n \times p} \approx \hat{Z}_{n \times p} = G_{n \times 2} H'_{2 \times p} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \\ \dots & \dots \\ g_{n1} & g_{n2} \end{pmatrix} \begin{pmatrix} h_{11} & h_{21} & \dots & h_{p1} \\ h_{12} & h_{22} & \dots & h_{p2} \end{pmatrix}$$

A seconda di come vengono specificate G e H , possiamo ottenere diverse rappresentazioni per i punti riga ed i punti colonna.

Rappresentazione A. dei punti riga e dei punti colonna

Le n righe della matrice dei dati (di dimensione p) possono essere rappresentate tramite la matrice $n \times 2$:

$$G = \sqrt{n-1} U_{(2)} = Z V_{(2)} \Gamma_{(2)}^{-1}$$

Questa matrice contiene gli scores standardizzati delle prime due componenti principali. In altri termini la matrice di varianze e covarianze degli scores normalizzati è la matrice identità di dimensione 2×2 :

$$\text{var}(\sqrt{n-1} U_{(2)}) = I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Le p colonne della matrice dei dati possono essere rappresentate tramite p frecce bidimensionali che partono dall'origine degli assi e che presentano

coordinate uguali a quelle delle p colonne della matrice che segue:

$$H' = \Gamma_{(2)} V'_{(2)} = \begin{pmatrix} \gamma_1 v_{11} & \gamma_1 v_{21} & \cdots & \gamma_1 v_{p1} \\ \gamma_2 v_{12} & \gamma_2 v_{22} & \cdots & \gamma_2 v_{p2} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} v_{11} & \sqrt{\lambda_1} v_{21} & \cdots & \sqrt{\lambda_1} v_{p1} \\ \sqrt{\lambda_2} v_{12} & \sqrt{\lambda_2} v_{22} & \cdots & \sqrt{\lambda_2} v_{p2} \end{pmatrix}$$

Le coordinate della j -esima freccia $(\sqrt{\lambda_1} v_{j1}, \sqrt{\lambda_1} v_{j2})$, $(j = 1, 2, \dots, p)$, non sono altro che i coefficienti di correlazione tra la j -esima variabile e le prime due componenti principali (matrice di componenti vista introdotta nella sezione precedente). La lunghezza della freccia in questo caso è esattamente uguale alla radice quadrata della comunaltà (ossia alla radice quadrata della quota di varianza della j -esima variabile spiegata dalle prime due componenti principali).

Il codice per ottenere la rappresentazione A è riportato di seguito.

```
%% A. BILOT CON CP STANDARDIZZATE E CORRELAZIONI
% i punti riga sono rappresentati dalle componenti principali standardizzate
% e i punti colonna tramite frecce la cui lunghezza rappresenta la correlazione
% tra le variabili originarie e le componenti principali;
close all
hold('on')
plot(sq1*U(:,1),sq1*U(:,2),'o')
text(sq1*U(:,1),sq1*U(:,2),Xtable.Properties.RowNames)

Vgam=V*Gamma;
zeroes = zeros(p,1);
quiver(zeroes,zeroes,Vgam(:,1),Vgam(:,2))
```

```

varlabs=Xtable.Properties.VariableNames;

dx=0.02;

dy=0.03;

text(Vgam(:,1)+dx,Vgam(:,2)+dy,varlabs,'Color','b');

% Vengono aggiunti gli assi cartesiani
xline(0)
yline(0)

```

L'output di questo codice è mostrato nella Figura 3.14.

Rappresentazione B dei punti riga e dei punti colonna

Le n righe della matrice dei dati (di dimensione p) possono essere rappresentate tramite la matrice $n \times 2$:

$$G = \sqrt{n-1}U_{(2)}\Gamma_{(2)} = ZV_{(2)}$$

In questo caso i punti riga sono gli scores non normalizzati cioè:

$$\text{var}(\sqrt{n-1}U_{(2)}\Gamma_{(2)}) = \text{var}(ZV_{(2)}) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

In questa rappresentazione i punti colonna (le frecce) non sono altro che le coordinate dei primi due autovettori:

$$H' = V'_{(2)} = \begin{pmatrix} v_{11} & v_{21} & \dots & v_{p1} \\ v_{12} & v_{22} & \dots & v_{p2} \end{pmatrix}$$

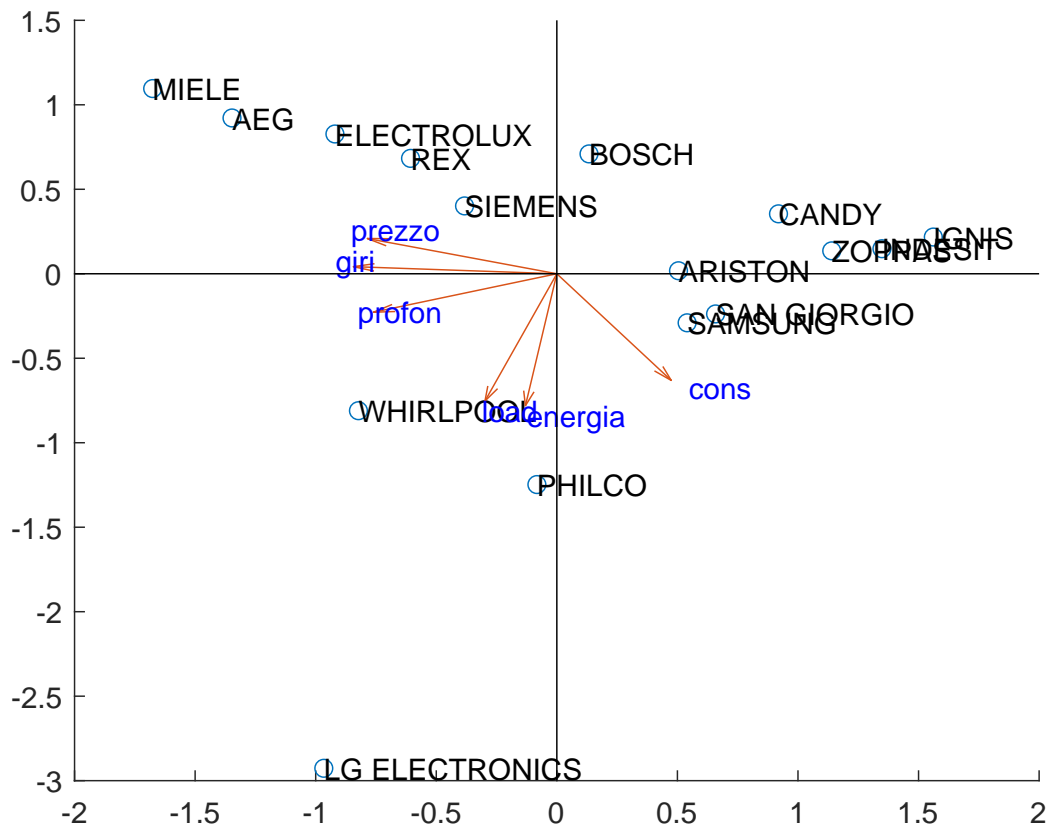


Figura 3.14: Biplot del dataset delle lavatrici. I punti riga sono rappresentati dalle componenti principali standardizzate e i punti colonna tramite frecce la cui lunghezza rappresenta la correlazione tra le variabili originarie e le componenti principali.

In questa rappresentazione, quindi, la lunghezza delle frecce non è uguale alla radice quadrata della comunalità ma è solo “funzione” della comunalità.

Il codice per ottenere la rappresentazione B è riportato di seguito.

```
% B. BILOT CON CP NON STANDARDIZZATE E AUTOVETTORI
```

```
% i punti riga sono rappresentati dalle componenti principali non standardizzate
```

```
% e i punti colonna tramite frecce la cui lunghezza rappresenta gli autovettori;
```

```
close all
```

```

hold('on')

plot(sqn1*U(:,1)*Gamma(1,1),sqn1*U(:,2)*Gamma(2,2),'o')

text(sqn1*U(:,1)*Gamma(1,1),sqn1*U(:,2)*Gamma(2,2),Xtable.Properties.RowNames

Vgam=V;

zeroes = zeros(p,1);

quiver(zeroes,zeroes,Vgam(:,1),Vgam(:,2))

varlabs=Xtable.Properties.VariableNames;

dx=0.02;

dy=0.03;

text(Vgam(:,1)+dx,Vgam(:,2)+dy,varlabs,'Color','b');

% Vengono aggiunti gli assi cartesiani

xline(0)

yline(0)

```

L'output di questo codice è mostrato nella Figura 3.15.

Ci chiediamo: è possibile avere una rappresentazione dinamica dei punti riga e dei punti colonna (freccie) in modo tale che le due rappresentazioni precedenti siano solo dei casi particolari? A questo scopo introduciamo due nuovi parametri ω , e α definiti nell'intervallo $[0, 1]$, e scriviamo la scomposizione in valori singolari come segue:

$$Z \approx \left[(\sqrt{n-1})^\omega U_{(2)} \Gamma_{(2)}^\alpha \right] \left[\Gamma_{(2)}^{1-\alpha} V'_{(2)} (\sqrt{n-1})^{1-\omega} \right]$$

In questo caso gli n punti riga possono essere rappresentati tramite la matrice

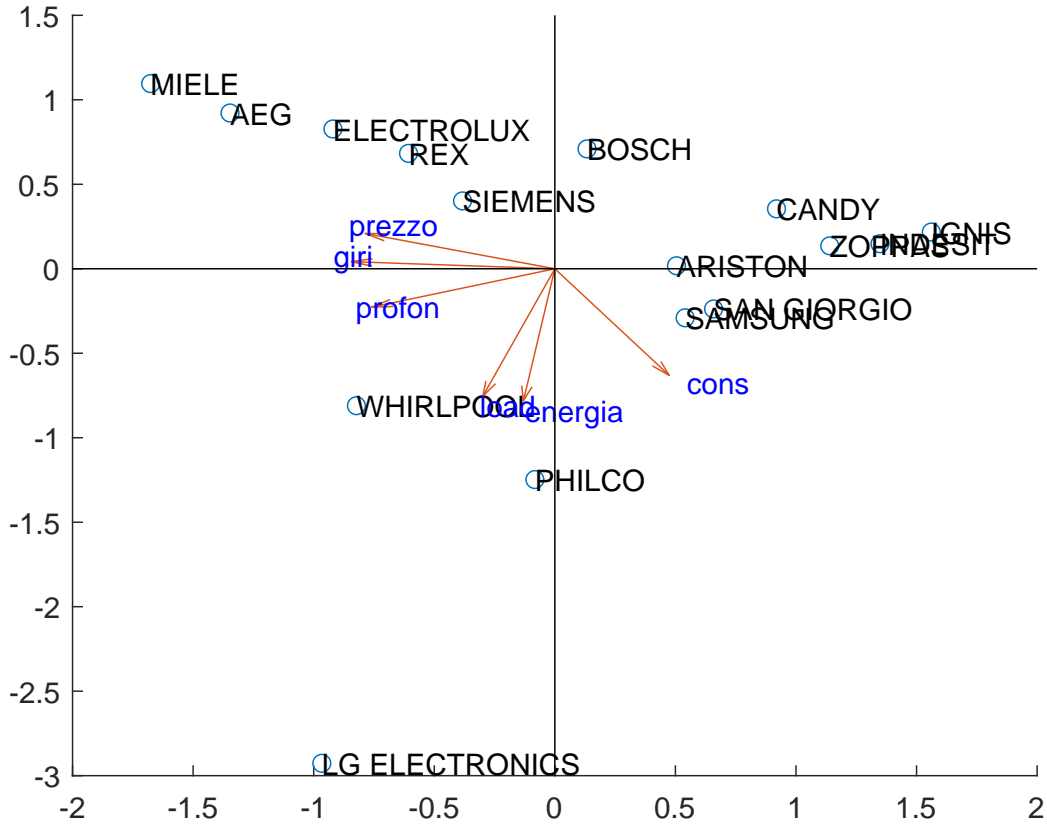


Figura 3.15: Biplot del dataset delle lavatrici. i punti riga sono rappresentati dalle componenti principali non standardizzate e i punti colonna tramite frecce la cui lunghezza coincide con i primi due autovettori.

$n \times 2$

$$\left[(\sqrt{n-1})^\omega U_{(2)} \Gamma_{(2)}^\alpha \right]$$

e i p punti colonna tramite la matrice $2 \times p$

$$[\Gamma_{(2)}^{1-\alpha} V'_{(2)} (\sqrt{n-1})^{1-\omega}]$$

La rappresentazione A vista in precedenza si ottiene quando $\omega = 1$ e $\alpha = 0$. Al contrario, la rappresentazione B si ottiene quando $\omega = 1$ e $\alpha = 1$. Ovviamente, tanto più ω diminuisce, tanto più la lunghezza delle frecce aumenta

e le coordinate dei punti riga si comprimono.

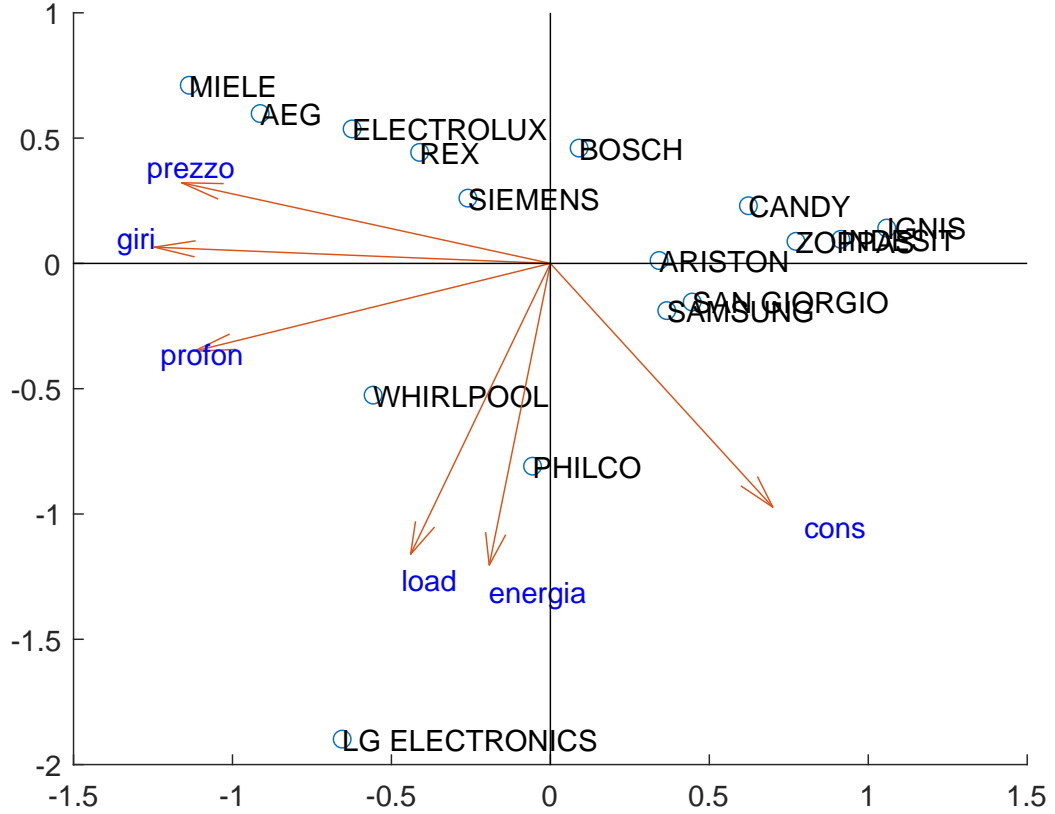


Figura 3.16: Biplot del dataset delle lavatrici. i punti riga sono rappresentati dalle coordinate della matrice $n \times 2 \sqrt{(n-1)^{0.6}} U_{(2)} \Gamma_{(2)}^{0.3}$ e i punti colonna tramite frecce le cui coordinate sono date dalla matrice $\sqrt{(n-1)^{1-0.6}} \Gamma_{(2)}^{1-0.3} V'_{(2)}$.

Il codice per ottenere questa rappresentazione quando $\omega = 0.6$ e $\alpha = 0.3$ è riportato di seguito.

```
close all
hold('on')
omega = 0.6;
alpha= 0.3;
```

```

PuntiRiga=sqn1^omega*U(:,1:2)*Gamma(1:2,1:2)^alpha;
PuntiColonna=V(:,1:2)*(Gamma(1:2,1:2)^(1-alpha))*sqn1^(1-omega);

plot(PuntiRiga(:,1),PuntiRiga(:,2),'o')
text(PuntiRiga(:,1),PuntiRiga(:,2),Xtable.Properties.RowNames)
zeroes = zeros(p,1);
quiver(zeroes,zeroes,PuntiColonna(:,1),PuntiColonna(:,2))
varlabs=Xtable.Properties.VariableNames;
dx=0.02;
dy=0.03;
text(PuntiColonna(:,1)+dx,PuntiColonna(:,2)+dy,varlabs,'Color','b');

% Vengono aggiunti gli assi cartesiani
xline(0)
yline(0)

```

L'output di questo codice è mostrato nella Figura 3.16.

Le dimensioni di \hat{Z} , $U_{(2)}$ e $V'_{(2)}$ sono riportate nella Figura 3.17. Le righe delle matrici $U_{(2)}$ sono le coordinate delle unità nel biplot e rappresentano i punti riga. Le p frecce che si dipartono dall'origine si riferiscono alle p colonne della matrice $V'_{(2)}$ e rappresentano i punti colonna. Le due frecce ad inversione che partano dalla matrice diagonale $\Gamma_{(2)}$ di dimensione 2×2 , che contiene i valori singolari, stanno ad indicare che questa matrice può essere inglobata nelle coordinate dei punti riga oppure nelle coordinate dei punti colonna, oppure in parte alle righe ed in parte alle colonne ($0 < \alpha < 1$).

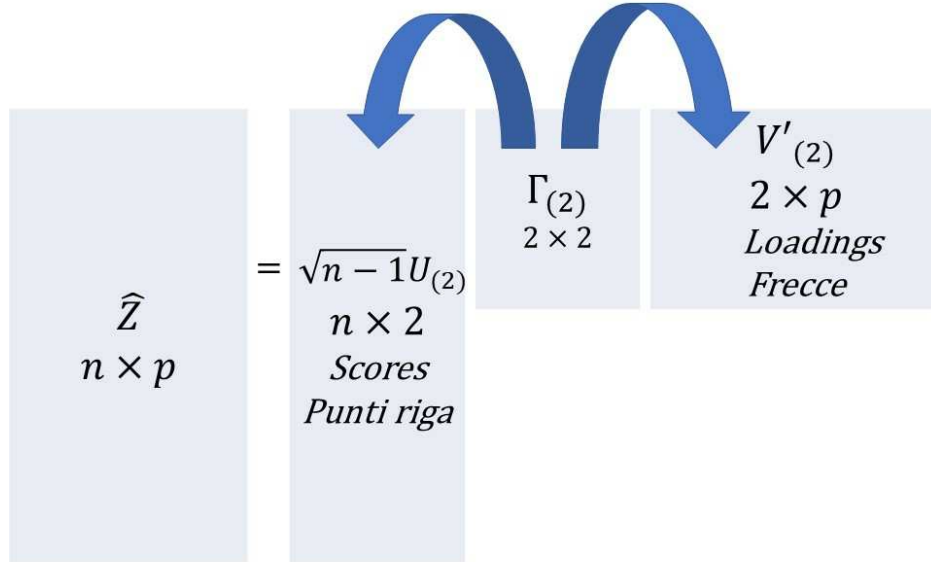


Figura 3.17: Rappresentazione grafica delle dimensioni delle 3 matrici della scomposizione in valori singolari quando si considerano solo i due valori singolari più grandi. Le due frecce ad inversione che partano dalla matrice diagonale $\Gamma_{(2)}$ di dimensione 2×2 che contiene i valori singolari, stanno ad indicare che questa matrice può essere inglobata nelle coordinate dei punti riga ($\alpha = 1$ rappresentazione B) oppure nelle coordinate dei punti colonna ($\alpha = 0$ rappresentazione A), oppure in parte nelle righe ed in parte nelle colonne ($0 < \alpha < 1$)

3.9 Componenti principali su \tilde{X} oppure su Z

Nelle sezioni precedenti abbiamo visto che l'analisi in componenti principali può partire dalla matrice \tilde{X} (scostamenti dalla media e gli autovalori sono quelli della matrice di covarianze) oppure dalla matrice Z (scostamenti standardizzati e gli autovalori sono quelli della matrice di correlazione). L'analisi in componenti principali con riferimento alla matrice di covarianze è correttamente applicabile quando le p variabili sono espresse nella stessa unità di

misura, presentano ordini di grandezza non molto diversi ed hanno anche variabilità non marcatamente differenti. Esempi salienti di situazioni di questo tipo sono:

- punteggi su scala da 0 a f ottenuti da n candidati in p batterie di test attitudinali.
- variabili espresse tutte in termini percentuali, ad esempio corrispondenti alla rilevazione in n comuni di una provincia delle percentuali di famiglie che posseggono p beni durevoli (lavastoviglie, barca, seconda auto, ...).
- numeri indici territoriali avendo posto il valore medio nazionale pari a 100 e considerando i valori che presentano le singole province italiane con riguardo a p indicatori dei livelli di ricchezza.

Dal punto di vista geometrico lavorare sulla matrice degli scostamenti standardizzati equivale a cercare il sottospazio che è più vicino alla nuova dei punti in termini di distanze Euclidee ponderate nello spazio originario. In altri termini, la scomposizione in svd della matrice Z

$$\begin{aligned} Z &= \sqrt{n-1}UV' \\ \tilde{X}D_c^{-1/2} &= \sqrt{n-1}UV' \end{aligned}$$

dove con D_c indichiamo la matrice diagonale che contiene sulla diagonale principale le varianze delle variabili originarie. $V_c = \text{diag}(\text{var}(X_1), \text{var}(X_2), \dots, \text{var}(X_p))$, corrisponde in termini di \tilde{X} ad avere una svd che chiamiamo “ponderata”.

$$\begin{aligned}\tilde{X} &= \sqrt{n-1} U \Gamma V' D_c^{1/2} \\ &= \sqrt{n-1} U \Gamma (D_c^{1/2} V)' \\ &= \sqrt{n-1} U \Gamma V^{*'}\end{aligned}$$

con $V^* = D_c^{1/2} V_c$ e $V^{*'} D_c^{-1} V^* = I_p$. In questo spazio metrico la distanza tra due vettori x_i e x_j è data da:

$$\|x_i - x_j\|_{D_c^{-1}}^2 = (x_i - x_j)' D_c^{-1} (x_i - x_j).$$

Il sottospazio cercato di dimensione k è quello che minimizza la seguente espressione:

$$\text{tr}((\tilde{X} - \hat{X})' D_c^{-1} (\tilde{X} - \hat{X})) = \sum_{i=1}^n \sum_{j=1}^p \frac{(\tilde{x}_{ij} - \hat{x}_{ij})^2}{\text{var}(X_j)} \quad (3.14)$$

$$= \sum_{i=1}^n (\tilde{x}_i - \hat{x}_i)' D_c^{-1} (\tilde{x}_i - \hat{x}_i) \quad (3.15)$$

fra tutte le matrici \hat{X} di rango al massimo uguale a k .

Quindi, se la matrice Z ha rango r nello spazio delle variabili standardizzate abbiamo la tradizionale svd:

$$Z = \sqrt{n-1} \sum_{j=1}^r \gamma_j u_j v_j'$$

Nello spazio delle variabili originarie in termini di scostamenti dalla media, implicitamente abbiamo la scomposizione in valori singolari ponderata come segue

$$\tilde{X} = \sqrt{n-1} \sum_{j=1}^r \gamma_j u_j v_j^{*'}$$

dove $V^* = (v_1^*, v_2^*, \dots, v_p^*)$ è tale per cui $V^{*'} D_c^{-1} V^* = I_p$ e $U'U = I_r$. Nello spazio di \tilde{X} questo corrisponde a minimizzare la quantità:

$$tr((\tilde{X} - \hat{X})' D_c^{-1} (\tilde{X} - \hat{X})) \quad (3.16)$$

fra tutte le matrici di rango k . La precedente quantità è minimizzata quando:

$$\hat{X} = \sqrt{n-1} \sum_{j=1}^k \gamma_j u_j v_j^{*'}$$

Questo si dimostra facilmente tenendo presente che l'equazione (3.16) può essere riscritta come:

$$tr \left(\left(\sum_{s=k+1}^p \gamma_s u_s v_s^{*'} \right)' D_c^{-1} \left(\sum_{j=k+1}^p \gamma_j u_j v_j^{*'} \right) \right)$$

con $V^{*'} D_c^{-1} V^* = I_p$ e $U'U = I_p$. Dato che $v_i^{*'} D_c^{-1} v_i^* = 1$ e $v_i^{*'} D_c^{-1} v_j^* = 0$ per $i \neq j$, si ha che

$$\begin{aligned} tr((\tilde{X} - \hat{X})' D_c^{-1} (\tilde{X} - \hat{X})) &= \sum_{j=k+1}^p \gamma_j^2 tr(u_j u_j') \\ &= \sum_{j=k+1}^p \gamma_j^2 \end{aligned}$$

Riassumendo, fare la scomposizione in valori singolari sulla matrice $Z =$

$\tilde{X}D_c^{-1/2}$, implicitamente significa andare a cercare un sottospazio Euclideo in cui la distanza tra i vettori nello spazio originario \tilde{X} è definita da una metrica euclidea ponderata ed andare a minimizzare l'equazione (3.15).

Similmente, se supponiamo di lavorare con unità statistiche che hanno diversa importanza dobbiamo associare un peso ad ogni riga della matrice dei dati. Sia D_r la matrice diagonale di dimensione $n \times n$ che contiene sulla diagonale principale questi pesi. $D_r = \text{diag}(w_1, w_2, \dots, w_n)$. Fare la scomposizione in valori singolari della matrice $D_r^{1/2}\tilde{X}D_c^{-1/2}$, matrice degli scostamenti standardizzati ponderati, equivale nello spazio originario definito da \tilde{X} andare a minimizzare

$$\text{tr}(D_r(\tilde{X} - \hat{X})'D_c^{-1}(\tilde{X} - \hat{X})) = \sum_{i=1}^n \sum_{j=1}^p w_i \frac{1}{\text{var}(X_j)} (\tilde{x}_{ij} - \hat{x}_{ij})^2$$

fra tutte le matrici \hat{X} di rango al massimo uguale a k .

Se cerchiamo la migliore matrice di rango due che rappresenta la matrice Z , se la matrice Z ha rango r , abbiamo che:

$$\begin{aligned} Z - \hat{Z} &= \sum_{j=3}^r \gamma_j u_j v_j' \\ D_r^{1/2} \tilde{X} D_c^{-1/2} - \hat{Z} &= \sum_{j=3}^r \gamma_j u_j v_j' \\ \tilde{X} - \hat{X} &= \sum_{j=3}^r \gamma_j u_j^* v_j^{*'} \end{aligned}$$

dove $V^* = (v_1^*, v_2^*, \dots, v_r^*)$ è tale per cui $V^{*'} D_c^{-1} V^* = I_p$ e $U^* = (u_1^*, u_2^*, \dots, u_r^*)$ è tale per cui $U^{*'} D_r U^* = I_r$. In altri termini, se ogni unità ha una diversa im-

portanza (definita da w_i) e ogni variabile ha una diversa variabilità (misurata da $1/\text{var}(X_j)$), fare la scomposizione in valori singolari di $Z = D_r^{1/2} \tilde{X} D_c^{-1/2}$ equivale ad andare a fare un'analisi in componenti principali generalizzata pesata, ossia minimizzare una somma dei quadrati ponderata in cui ogni dimensione ed ogni unità presenta un peso diverso. La variabilità totale della matrice Z di dimensione $n \times p$ di rango r è data da

$$\text{tr} \left((D_r^{1/2} \tilde{X} D_c^{-1/2})' (D_r^{1/2} \tilde{X} D_c^{-1/2}) \right) = \sum_{i=1}^n w_i \tilde{x}_i' D_c^{-1} \tilde{x}_i = \sum_{j=1}^r \gamma_j^2 = p \sum_{i=1}^n w_i \quad (3.17)$$

dove i γ_j sono i valori singolari della matrice $D_r^{1/2} \tilde{X} D_c^{-1/2}$.

Dimostrazione

L'equazione (3.17) non è altro che una somma pesata delle distanze di Mahalanobis (v. equazione 2.9).

$$\begin{aligned} &= \sum_{i=1}^n w_i \tilde{x}_i' D_c^{-1} \tilde{x}_i \\ &= \text{tr} \left(\sum_{i=1}^n w_i D_c^{-1} \tilde{x}_i \tilde{x}_i' \right) \\ &= \text{tr} \left(D_c^{-1} \sum_{i=1}^n w_i \tilde{x}_i \tilde{x}_i' \right) \\ &= \sum_{i=1}^n w_i \text{tr} (D_c^{-1} S) \\ &= \sum_{i=1}^n w_i \times p \end{aligned}$$

Esercizio: generare una matrice di dimensione 50×10 di rango $r = 4$. Le prime tre colonne contengono numeri da $N(0, 1)$. Generare un vettore di pesi

w da $U(0,1)$. Controllare che la matrice X abbia le dimensioni ed il rango richiesto. Creare la matrice Z degli scostamenti standardizzati ponderati. Verificare che

$$\text{tr}(Z'Z) = \sum_{i=1}^{50} \sum_{j=1}^{10} z_{ij}^2 = \sum_{l=1}^4 \gamma_l^2 = 4 \sum_{i=1}^n w_i$$

e che le precedenti espressioni siano uguali all'equazione (3.17).

Soluzione.

```

n=50;

p1=3;

X=[randn(n,p1) ones(n,7)];

p=size(X,2);

r=rank(X);

disp(["Dimensioni X" string(n) "x" string(p)])

disp(["Rango della matrice X=" string(r)])

% w = vettore dei pesi

w=rand(n,1);

% meX = vettore 1xp che contiene le medie pesate

meX=sum(X.*w);

% Dr = matrice diag con pesi delle righe sulla

% diagonale

Dr=diag(w);

% Xtilde = matrice degli scostamenti dalla media

Xtilde=X-meX;

% S = matrice di covarianze

S=Xtilde'*Dr*Xtilde/sum(w);

```

```

% Dc = matrice diagonale con Var(X_j) sulla diagonale
Dc=diag(diag(S));
% Dcminus1 inversa di Dc
Dcminus1=Dc^-(1);
% totsum conterrà la somma pesata delle distanze di Mahalanobis
% al quadrato
totsum=0;
for i=1:n
    xitilde=Xtilde(i,:);
    totsum=totsum+w(i)*xitilde'*Dcminus1*xitilde;
end
% Z = matrice degli scostamenti standardizzati ponderati
Z=Dr^(1/2)*Xtilde*Dc^(-1/2);
disp(["tr(Z'Z)=" string(trace(Z'*Z))]);
disp(["sum z_ij^2 " string(sum(Z.^2,'all'))])
% svd di Z
[U,Gamma,V]=svd(Z,'econ');
% somma dei quadrati dei valori singolari
% La matrice Gamma ha 4 valori singolari uguali a 0
sumgamsquared=sum(Gamma.^2,'all');
disp(["sum \gam_i^2=" sumgamsquared])
disp(["r*sum(w)=" string(p*sum(w))])

```

Nel capitolo che segue andiamo ad applicare i concetti appena vista sulla scomposizione in valori singolari ponderata, per rappresentare in un

sottospazio a dimensione ridotta l'informazione contenuta nella tabella di contingenza.

Capitolo 4

L'analisi delle corrispondenze

L'analisi delle corrispondenze è una tecnica statistica multivariata per visualizzare e descrivere le associazioni fra due o più variabili qualitative le cui modalità sono state classificate in una tabella di contingenza a due o più vie. Quando la tabella di contingenza di partenza ha dimensioni superiori a 2×2 , risulta difficile una rappresentazione grafica diretta che ponga in luce le relazioni tra le diverse modalità che presentano le due variabili in esame. Una volta accertato che il valore del test χ^2 è significativo, si pone il problema di capire quali sono le combinazioni di righe e colonne nella tabella di contingenza che presentano la relazione più forte.

Il procedimento che sta alla base del metodo consiste nel “geometrizzare il problema”, nel senso che le righe e le colonne della tabella di contingenza, opportunamente ricodificate, vengono intese come punti geometrici in due diversi spazi multidimensionali, nei quali è definita una distanza, dando vita quindi a due “nuvole” di punti. Per poterne decifrare la struttura, ciascuna nuvola viene proiettata in un sottospazio a due dimensioni (un piano). Nella

stessa maniera delle componenti principali, i sottospazi sono scelti in maniera ottimale, in modo tale che i punti proiettati diano una rappresentazione il più possibile fedele, della nube originaria. Grazie alle preventive trasformazioni operate simmetricamente sulle righe e sulle colonne della matrice dei dati, è possibile far coincidere i due piani, ottenendo così una rappresentazione grafica unica sulla quale le righe e le colonne della matrice vengono ad essere rappresentate dalle proiezioni dei loro punti rappresentativi. L'interpretazione delle prossimità tra proiezioni sulla mappa conduce l'analista a risalire alle prossimità tra punti delle nuvole nel loro spazio multidimensionale e perciò a riconoscere i legami tra le caratteristiche il cui l'insieme dei dati è ripartito.

La metodologia dell'analisi delle corrispondenze è stata introdotta dalla scuola francese guidata da J.P. Benzecri all'inizio degli anni sessanta. Quando l'analisi delle corrispondenze è relativa a due variabili si parla di analisi delle corrispondenze semplici. Qualora invece si considerano congiuntamente più di due variabili che danno luogo a tabelle di contingenza a tre o più vie, si parla di analisi delle corrispondenze multiple.

L'analisi delle corrispondenze è impiegata in contesti molto svariati. Ad esempio, dal punto di vista di marketing, l'analisi della corrispondenza risponde a domande come:

1. Ci sono lacune nel mercato che potrebbero essere colmate da un determinato business?
2. Il posizionamento del marchio è corretto?

3. L'azienda potrebbe differenziarsi dalla concorrenza?
4. Quali attributi possiedono i concorrenti o, in alternativa, possiedono una determinata attività?
5. Su quali fasce di età è preferibile impostare una campagna pubblicitaria?

Dal punto di vista medico l'analisi delle corrispondenze consente di sia di comprendere meglio quali sono le cause che portano ad un peggioramento della condizione fisica, sia di capire immediatamente quali caratteristiche hanno le persone che presentano determinati sintomi.

I punti di forza nell'analisi delle corrispondenze consistono nel fatto che questa tecnica, senza fare ipotesi distributive, consente di capire in maniera immediata, tramite metodi di visualizzazione semplici e potenti, quali righe delle tabelle sono maggiormente associate a quali colonne.

Consideriamo ora una serie di esempi in cui coppie di fenomeni qualitativi sono stati classificati in tabelle di contingenze.

La Tabella 4.1 riporta la posizione sulla scienza ed il grado di scolarizzazione di 871 individui. In questo caso risulta importante capire quali sono le categorie di scolarizzazione più (meno) favorevoli alla scienza.

La Tabella 4.2 riporta la professione ed il titolo di studio di 426 intervistati. In questo caso l'obiettivo è capire quali sono le professioni tipiche per determinati titoli di studio.

Tabella 4.1: Tabella di contingenza tra il grado di scolarizzazione e la posizione verso la scienza per un campione di 871 persone. 1= Licenza elementare, 2= Licenza media, 3=Diploma di scuola media superiore, 4=Laurea triennale, 5=Laurea specialistica, 6=Dottorato di Ricerca

	1	2	3	4	5	6
Per Niente Favorevole	6	34	19	6	4	2
Poco Favorevole	10	93	47	12	5	7
Indifferente	11	95	55	18	11	15
Favorevole	7	112	82	37	16	27
Molto Favorevole	4	44	39	21	13	19

Tabella 4.2: Tabella di contingenza tra il grado di istruzione e la professione lavorativa in un campione di 426 persone.

	Elementare	Laurea	Media	Superiore
Altro	2	2	0	3
Artigiano	0	0	3	1
Casalinga	44	0	28	10
Commerciante	1	0	2	2
Dirigente	0	3	1	0
Disoccupato	1	2	2	0
Impiegato	0	11	12	40
Imprenditore	0	2	1	0
Insegnante	1	13	1	5
Libero Profess.	0	7	1	7
Operaio	8	0	31	22
Pensionato	70	2	28	27
Studente	0	9	1	20

La Tabella 4.3 riporta la tabella di contingenza riferita alla posizione sulla pena di morte ed al partito politico di appartenenza in un'indagine effettuata su 943 cittadini Americani. In questo caso è interessante capire quali sono i partiti più (meno) favorevoli alla pena capitale.

Tabella 4.3: Tabella di contingenza relativa all'appartenenza al partito politico e alla posizione sulla pena di morte in un campione di 943 cittadini Americani

	CONTRARIO	FAVOREVOLE	Non So
REPUBBLICANO	21	141	9
DEMOCRATICO	53	85	3
ATTIVISTA			
INDIPENDENTE	17	69	16
DEMOCRATICO	49	137	13
REPUBBLICANO	9	99	6
ATTIVISTA			
DEMOCRATICO	27	89	7
SIMPATIZZANTE			
REPUBBLICANO	12	75	6
SIMPATIZZANTE			

La Tabella 4.4 mostra le occorrenze dei 28 membri dell'UE nel flusso commerciale di capi non provenienti dall'Unione Europea, secondo cinque fasce di prezzo. x_1 denota il segmento di prezzo più basso e x_5 il segmento di prezzo più alto. In questo caso l'obiettivo è capire se qualche nazione è legata in maniera sistematica ad un livello di prezzo più basso in modo da creare una sottofatturazione per pagare ridotti dazi doganali.

La Tabella 4.5 riporta la tipologia di marca di dentifricio utilizzata in 4 regioni italiane. In questo caso, l'obiettivo è cercare di comprendere in quale regione si acquista prevalentemente il dentifricio di marca commerciale, per capire in quale territorio il fattore prezzo risulta più marcato.

Tabella 4.4: Dati sull'importazione di generi di abbigliamento nella UE. Tabella di contingenza tra i 28 stati membri dell'Unione Europea (dati raccolti prima della Brexit) e 5 fasce di prezzo. La tabella di contingenza contiene le occorrenze dei flussi commerciali del paese, per un ampio insieme di tipologie di vestiti: x_1 denota il segmento di prezzo più basso e x_5 il segmento di prezzo più alto. In tutto ci sono 4373 flussi commerciali

	x_1	x_2	x_3	x_4	x_5
GB	134	76	43	50	49
SK	173	62	20	23	16
BG	67	76	48	36	23
IE	11	21	31	36	52
BE	25	32	57	60	58
ES	32	42	40	67	67
PL	20	35	31	41	41
FI	10	16	23	23	24
GR	54	28	29	30	23
HU	12	19	14	15	20
SI	9	10	14	20	23
NL	52	43	38	47	54
IT	21	36	33	30	36
RO	85	74	55	31	22
AT	3	8	12	12	25
FR	28	33	40	31	45
HR	9	17	23	19	34
SE	18	36	44	35	40
CZ	12	24	22	25	37
DK	16	32	35	39	38
DE	28	39	36	41	54
LT	3	15	22	25	24
PT	30	40	28	20	26
EE	8	10	12	13	17
LU	2	1	2	3	3
MT	29	10	16	8	9
LV	47	51	29	19	12
CY	7	19	20	26	9

Tabella 4.5: Tipologia di marca di dentifricio prevalentemente utilizzata da 1576 consumatori appartenenti a 4 regioni Italiane. (A=marca commerciale, B=Marca industriale, C=Indifferente). L'ultima riga e l'ultima colonna contengono rispettivamente le frequenze marginali delle righe e delle colonne

	Liguria	Lombardia	Piemonte	Veneto	Totale $n_{i.}$
A	49	111	13	49	222
B	16	551	241	7	815
C	34	358	30	117	539
Totale $n_{.j}$	99	1020	284	173	1576

4.1 Notazione

Per presentare i concetti di base dell'analisi delle corrispondenze semplice, si fa riferimento ad una generica tabella con I righe e J colonne. Utilizzando la simbologia già vista nel capitolo sull'associazione si indicano rispettivamente con n_{ij} e $f_{ij} = n_{ij}/n$ le frequenze assolute e relative della tabella dove $\sum_{i=1}^I \sum_{j=1}^J n_{ij} = n$ è la numerosità totale ed il totale delle frequenze relative soddisfa il vincolo di somma unitaria cioè $\sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$. Le matrici contenenti rispettivamente le frequenze assolute e relative sono indicate con i simboli N e P . Utilizzando la notazione matriciale abbiamo che $(1_{I \times 1})' N 1_{J \times 1} = n$ e $(1_{I \times 1})' P 1_{J \times 1} = 1$ dove $1_{I \times 1}$ e $1_{J \times 1}$ sono due vettori colonna di lunghezza pari a I e J con elementi tutti uguali ad 1. Nel linguaggio dell'analisi delle corrispondenze, la frequenza marginale relativa della riga i -esima ($f_{i.} = n_{i.}/n$) viene detta massa della riga i -esima. Il vettore colonna

di lunghezza I che contiene le masse di riga viene indicato con r .

$$r = \begin{pmatrix} f_{1.} \\ f_{2.} \\ \dots \\ f_{I.} \end{pmatrix}$$

In termini matriciali $r = P1_J$. Indichiamo infine con D_r la matrice diagonale di dimensione $I \times I$ che contiene sulla diagonale principale le masse di riga: $D_r = \text{diag}(f_{1.}, f_{2.}, \dots, f_{I.})$. Similmente, la frequenza marginale relativa della colonna j -esima ($f_{.j} = n_{.j}/n$) viene detta massa della colonna j -esima. Il vettore colonna di lunghezza J che contiene le masse di colonna viene indicato con c .

$$c' = \begin{pmatrix} f_{.1} & f_{.2} & \dots & f_{.J} \end{pmatrix}$$

In termini matriciali $c = P'1_I$. Indichiamo infine con D_c la matrice diagonale di dimensione $J \times J$ che contiene sulla diagonale principale le masse di colonna: $D_c = \text{diag}(f_{.1}, f_{.2}, \dots, f_{.J})$.

A questo punto definiamo le tabelle dei profili dividendo il valore di ogni elemento della matrice N (oppure della matrice P) per il corrispondente totale o di riga o di colonna.

Definizione: si dice profilo riga i -esimo (o profilo della i -esima riga) e si indica con r_i il vettore:

$$r'_i = \begin{pmatrix} f_{i1}/f_{i.} & f_{i2}/f_{i.} & \dots & f_{iJ}/f_{i.} \end{pmatrix} = \begin{pmatrix} n_{i1}/n_{i.} & n_{i2}/n_{i.} & \dots & n_{iJ}/n_{i.} \end{pmatrix}$$

La matrice dei profili riga di dimensione $I \times J$ viene indicata con R^1 .

$$R = \begin{pmatrix} r'_1 \\ r'_2 \\ \dots \\ r'_I \end{pmatrix}$$

In termini matriciali $R = D_r^{-1}P$. La tabella 4.6 riporta la matrice dei profili riga associati alla Tabella 4.5.

Tabella 4.6: Matrice dei profili riga ($R_{3 \times 4}$) della Tabella 4.5. In questa matrice la somma di ogni riga è pari a 1. L'ultima riga e l'ultima colonna di questa tabella contengono rispettivamente le masse di colonna (c') e di riga (r)

	Liguria	Lombardia	Piemonte	Veneto	Massa di riga $r_{3 \times 1} =$
A	0.221	0.500	0.059	0.221	0.141 f_1 .
B	0.020	0.676	0.296	0.009	0.517 f_2 .
C	0.063	0.664	0.056	0.217	0.342 f_3 .
Massa di colonna $c'_{1 \times 4} =$	0.063	0.647	0.180	0.110	1
	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	

Definizione: si dice profilo colonna j -esimo (o profilo della j -esima colon-

¹Questo con un piccolo abuso di notazione in quanto negli altri capitoli avevamo utilizzato il simbolo R per indicare la matrice di correlazione della matrice dei dati X di variabili quantitative.

na) e si indica con c_j il vettore di lunghezza I che segue:

$$c_j = \begin{pmatrix} f_{1j}/f_{.j} \\ f_{2j}/f_{.j} \\ \dots \\ f_{Ij}/f_{.j} \end{pmatrix} = \begin{pmatrix} n_{1j}/n_{.j} \\ n_{2j}/n_{.j} \\ \dots \\ n_{Ij}/n_{.j} \end{pmatrix}$$

La matrice dei profili colonna di dimensione $J \times I$ viene indicata con C

$$C = \begin{pmatrix} c'_1 \\ c'_2 \\ \dots \\ c'_J \end{pmatrix}$$

In termini matriciali $C = D_c^{-1}P'$.

La tabella 4.7 riporta la matrice trasposta dei profili colonna associati alla Tabella 4.5.

Tabella 4.7: Matrice trasposta dei profili colonna ($C_{4 \times 3}$) della Tabella 4.5. In questa matrice la somma di ogni colonna è pari a 1. L'ultima riga e l'ultima colonna di questa tabella contengono rispettivamente le masse di colonna (c') e di riga (r)

	Liguria	Lombardia	Piemonte	Veneto	Massa di riga $r_{3 \times 1} =$
A	0.495	0.109	0.046	0.283	0.141 f_1 .
B	0.162	0.540	0.849	0.040	0.517 f_2 .
C	0.343	0.351	0.106	0.676	0.342 f_3 .
Massa di colonna $c'_{1 \times 4} =$	0.063	0.647	0.180	0.110	1
	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	

Esercizio.

Data la tabella di contingenza riportata nella Tabella 4.5, creare la matrice di corrispondenze P (*correspondence matrix*). Verificare tramite moltiplicazioni matriciali che la somma degli elementi di P è pari a 1. Calcolare le matrici dei profili riga e la trasposta dei profili colonna. Chiamare la matrice R dei profili riga `ProfilesRows` e la trasposta dei profili colonna (C') `ProfilesCols`. Calcolare i vettori r e c che contengono rispettivamente le masse di riga e di colonna. Creare le matrici D_r e D_c .

Soluzione

```
N=[49 111 13 49
16 551 241 7
34 358 30 117];

% I = numero di righe della tabella di contingenza
% J = numero di colonne della tabella di contingenza
[I,J]=size(N);

%% Calcolare la matrice delle corrispondenze P che contiene le frequenze
% relative.

% n= numero di unità del campione
n=sum(N,'all');

% P = (matrice di corrispondenza =correspondence matrix) contiene le
% frequenze relative f_ij
P = (1/n) * N;
```

```

% Verifico tramite moltiplicazione matriciale che la somma degli elementi
% di P è pari 1.
onesI1=ones(I,1);
onesJ1=ones(J,1);
sumelP=onesI1'*P*onesJ1;
assert(abs(sumelP-1)<1e-12,"La somma degli elementi di P non è 1")

%% Calcolo delle matrici dei profili riga e colonna

% ProfilesRows = matrice che contiene i profili riga
% Si divide ogni riga per il suo totale
% ProfilesRows ha dimensione IxJ
ProfilesRows = N./sum(N,2);

% ProfilesCols = matrice trasposta dei profili colonna
% Si divide ogni colonna per il suo totale e si fa la trasposta
% La matrice dei profili colonna ha dimensione JxI
% ProfilesCols ha dimensione IxJ
ProfilesCols = (N./sum(N,1));

%% Calcolo vettori r e c
% r= vettore che contiene le masse di righe
% = centroidi dei profili colonna

```

```

r=sum(N,2)/n;

% c= vettore che contiene le masse di colonna
% = centroidi dei profili di riga
c=(sum(N,1)/n)';

%% Costruzione matrici Dr e Dc
% Queste matrici diagonali contengono rispettivamente i profili
% medi di colonna e riga sulla diagonale principale.
Dr = diag(r);
Dc = diag(c);

```

Dall'esame delle Tabelle 4.6 e 4.7 emerge che nel campione degli intervistati il consumo maggiore è per la marca *B* (punto con la più alta massa di riga) e che nell'ambito delle regioni, il numero più elevato degli intervistati risiede in Lombardia (punto colonna con la massa più elevata). Non è ben chiaro, però, qual è la relazione tra le righe e colonne della tabella di contingenza. Vedremo che tramite l'analisi delle corrispondenze si riesce in maniera semplice a mostrare la complessità delle relazioni presenti in una tabella di contingenza anche di dimensioni piuttosto elevate.

Gli I profili riga possono essere visti come I punti in uno spazio di dimensione $J - 1$. Dato che la somma degli elementi di ogni profilo riga è pari a 1, l'ultimo elemento di ogni vettore può essere ottenuto come combinazione lineare degli altri elementi di conseguenza la matrice degli I profili riga può avere al massimo rango $J - 1$. Similmente i J profili colonna possono essere visti come J punti in uno spazio di dimensione $I - 1$. Entrambi gli spazi

sono ponderati nel senso che il peso di ciascuno degli I punti è dato dall'importanza della riga i ossia dalla sua massa ($f_{i.}$) e la distanza tra due generici punti riga avviene con la metrica χ^2 , ossia con i pesi definiti dall'inversa delle masse di colonna. In altri termini, la distanza (al quadrato) tra due generici profili riga r_i e r_{i^*} è data dalla seguente metrica Euclidea ponderata

$${}_w d_{ij}^2 = (r_i - r_{i^*})' D_c^{-1} (r_i - r_{i^*}) = \sum_{j=1}^J \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i^*j}}{n_{i^*.}} \right)^2. \quad (4.1)$$

Utilizzando la distanza Euclidea non ponderata si attribuirebbe un peso uguale a tutte le colonne delle tabella di contingenza. Questo aspetto è indesiderabile qualora nella tabella di contingenza si desidera dare un peso rilevante alle colonne in cui vi è una grande distanza tra due punti riga, ma il cui totale di colonna è piccolo. In tale circostanza, infatti, l'eventuale grande distanza tra due punti riga verrebbe mascherata, dal momento che l'impatto della colonna in considerazione risulta trascurabile, quando confrontata con le rimanenti colonne. Moltiplicando ogni addendo della somma nell'espressione (4.1) per $n/n_{.j}$ si evita questo inconveniente. La metrica (distanza) riportata nell'equazione (4.1) è nota in letteratura con il nome di distanza del χ^2 e spesso per indicarla si usa il simbolo $\chi^2 d_{ij}^2$.

Una proprietà aggiuntiva della metrica χ^2 è la seguente: se due profili riga uguali o proporzionali vengono aggregati in un unico profilo riga con massa pari alla somma delle masse, la configurazione dei punti non cambia né si modificano le distanze tra i profili colonna. Ovviamente la stessa proprietà vale per i profili colonna. Questa proprietà è importante poiché consente di raggruppare due o più righe (colonne) proporzionali in una sola, ridu-

cendo le dimensioni dello spazio di riferimento garantendo l'invariabilità dei risultati indipendentemente da come le variabili sono state originariamente codificate².

Dato che i profili riga hanno una diversa importanza (diversa massa), è naturale trovare il centroide dei profili riga tramite una media aritmetica ponderata degli stessi. È facile constatare che il profilo medio di riga per la colonna j -esima è ottenibile come media ponderata delle frequenze relative $f_{ij}/f_{i.}$ (distribuzione parziale di riga riferita alla colonna j) con pesi uguali alle masse di riga. La seguente media ponderata

$$\sum_{i=1}^I \frac{f_{ij}}{f_{i.}} f_{i.} = \sum_{i=1}^I f_{ij} = f_{.j} \quad j = 1, 2, \dots, J$$

non è altro che la frequenza marginale della colonna j -esima (massa di colonna j -esima). Il vettore dei profili medi di riga, quindi, è esattamente uguale a c . Utilizzando le moltiplicazioni matriciali³

$$c = R'r$$

Similmente, la frequenza relativa marginale della riga i -esima (massa della riga i) non è altro che la media ponderata delle $f_{ij}/f_{.j}$ con pesi uguali alle

²Questa proprietà è nota in letteratura con il nome di equivalenza distributiva.

³ Il vettore delle medie ponderate dei profili riga è dato da $R'r/r'1_I$. Tenendo presente che $r'1_{I \times 1} = 1$ e che $D_r^{-1}r = 1_{I \times 1}$ si ha che

$$R'r/r'1_I = R'r = (D_r^{-1}P)'r = P'D_r^{-1}r = P'1_{I \times 1} = c$$

masse di colonna.

$$\sum_{j=1}^J \frac{f_{ij}}{f_{.j}} f_{.j} = \sum_{j=1}^J f_{ij} = f_{i.} \quad i = 1, 2, \dots, I$$

Il vettore dei profili medi di colonna, quindi, non è altro che il vettore che r che contiene le masse di riga. Utilizzando le moltiplicazioni matriciali $r = C'c$;

La distanza (al quadrato) di ogni punto riga dal centroide nella metrica χ^2 è data da:

$$\chi^2 d_{ic}^2 = (r_i - c)' D_c^{-1} (r_i - c) = \sum_{j=1}^J \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \quad (4.2)$$

La Tabella 4.8 riporta le distanze dei diversi punti riga e colonna e le relative masse. Da questa tabella emerge che il profilo riga più distante è relativo alla marca industriale e che la regione con il profilo colonna più distante è la Liguria.

Tabella 4.8: Distanze al quadrato di ogni profilo riga dal profilo medio $\chi^2 d_{ic}^2$ e masse di riga (r). Distanze al quadrato di ogni profilo colonna dal profilo medio $\chi^2 d_{jr}^2$ e masse di colonna (c).

	$\chi^2 d_{ic}^2$	r		$\chi^2 d_{jr}^2$	c
$\chi^2 d_{1c}^2$	0.6247	0.1409	$\chi^2 d_{1r}^2$	1.1345	0.0628
$\chi^2 d_{2c}^2$	0.1983	0.5171	$\chi^2 d_{2r}^2$	0.0086	0.6472
$\chi^2 d_{3c}^2$	0.1914	0.342	$\chi^2 d_{3r}^2$	0.44	0.1802
			$\chi^2 d_{4r}^2$	0.91	0.1098

Come la matrice degli scostamenti dalla media \tilde{X} si può scrivere come $\tilde{X} = X - 1_{n \times 1} \bar{x}'$, la matrice degli scostamenti dalla media dei profili riga (che

denotiamo con \tilde{R} si può scrivere come

$$\tilde{R} = R - 1_{I \times 1} c'$$

Nel caso della matrice dei dati X , la metrica ponderata per ottenere gli scostamenti standardizzati era basata su $D_c = (\text{var}(X_1), \text{var}(X_2), \dots, \text{var}(X_p))$ con la metrica del χ^2 l'analoga della matrice Z al caso in cui abbiamo i pesi delle righe e delle colonne si può scrivere come

$$D_r^{1/2} \tilde{R} D_c^{-1/2} = D_r^{1/2} (R - 1_{I \times 1} c') D_c^{-1/2} \quad (4.3)$$

La variabilità totale di questa matrice (che nel linguaggio dell'analisi delle corrispondenze si chiama inerzia totale dei punti riga ed è spesso denotata con $in(I)$) è data da

$$= \text{tr} \left((D_r^{1/2} (R - 1_{I \times 1} c') D_c^{-1})' (D_r^{1/2} (R - 1_{I \times 1} c') D_c^{-1}) \right) \quad (4.4)$$

$$= \sum_{i=1}^I f_i (r_i - c)' D_c^{-1} (r_i - c) \quad (4.5)$$

Il nostro obiettivo è, come al solito, proiettare i punti riga in un sottospazio di dimensione ridotta in modo da perdere il meno possibile informazione. L'analisi delle corrispondenze, quindi non è altro che un'analisi in componenti principali ponderata. Il peso di ogni punto riga è determinato da $D_r^{1/2}$ ed ogni colonna ha un'importanza misurata da $D_c^{-1/2}$.

Il sottospazio migliore di dimensione k che minimizza la somma dei qua-

drati degli scostamenti tra valori osservati e valori adattati \hat{Z}

$$tr((D_r^{1/2}(R - 1_{I \times 1}c')D_c^{-1} - \hat{Z})(D_r^{1/2}(R - 1_{I \times 1}c')D_c^{-1} - \hat{Z}))$$

è dato da:

$$Z_{(k)} = U_{(k)}\Gamma_{(k)}V'_{(k)} = \sum_{j=1}^k u_i \gamma_i v'_i$$

dove la sommatoria sia estende ai primi k valori singolari della matrice $D_r^{1/2}(R - 1_{I \times 1}c')D_c^{-1}$. Tenendo presente che

$$Z = D_r^{1/2}(R - 1_{I \times 1}c')D_c^{-1/2} = U\Gamma V'$$

segue immediatamente che

$$R - 1_{I \times 1}c' = D_r^{-1/2}U\Gamma(D_c^{1/2}V)' \quad (4.6)$$

La matrice $D_c^{1/2}V$ contiene i nuovi assi cartesiani e la matrice $D_r^{-1/2}U\Gamma$ contiene le coordinate dei punti proiettati nel nuovo sistema di assi cartesiani (v. equazione (3.13)). La matrice Z può essere riscritta come:

$$\begin{aligned} Z &= D_r^{1/2}(R - 1_{I \times 1}c')D_c^{-1/2} \\ &= D_r^{-1/2}D_r(R - 1_{I \times 1}c')D_c^{-1/2} \\ &= D_r^{-1/2}(D_r R - D_r 1_{I \times 1}c')D_c^{-1/2} \\ &= D_r^{-1/2}(P - rc')D_c^{-1/2} \end{aligned} \quad (4.7)$$

Quest'ultima formulazione è quella più utilizzata, perché mostra che l'a-

nalisi è simmetrica nel senso che se partiamo dai punti colonna otteniamo gli stessi risultati. Più precisamente, dato che i J profili colonna presentano un centroide definito da r (il vettore delle masse di riga), e pesi dati dalla matrice $D_c^{1/2}$ (vettore c) e metrica χ^2 definita dall'inversa dagli elementi di r , la matrice standardizzata dei profili colonna è

$$D_c^{1/2}(C - 1_{J \times 1}r')D_r^{-1/2}$$

Questa matrice può essere riscritta come:

$$D_c^{-1/2}(D_c C - D_c 1_{J \times 1}r')D_r^{-1/2} = D_c^{-1/2}(P' - cr')D_r^{-1/2}$$

Questa matrice non è altro che la trasposta della matrice dei profili riga standardizzati (v. equazione 4.3). Dato che $tr(X'X) = tr(XX')$ (v. equazione (1.9)), segue immediatamente che l'inerzia totale dei punti riga è esattamente uguale all'inerzia dei punti colonna. Quindi, la somma dei quadrati delle distanze dei profili riga dal profilo medio ponderate con le masse di righe (equazione 4.5) è esattamente uguale alla somma dei quadrati delle distanze dei profili colonna dal rispettivo profilo medio ponderate con le masse di colonna:

$$in(J) = \sum_{j=1}^J f_{.j}(c_j - r)'D_r^{-1}(c_j - r)$$

A sua volta questa somma ponderata delle distanze coincide con la somma dei quadrati dei valori singolari $\sum_{s=1}^{\min(I-1, J-1)} \gamma_s^2$ della matrice (4.7). È interessante osservare che il valore dell'inerzia coincide con il valore dell'indice

χ^2/n introdotto nel capitolo sull'associazione.

$$in(I) = in(J) = \frac{\chi^2}{n}$$

Infatti

$$in(I) = in(J) = \sum_{i=1}^I \sum_{j=1}^J \frac{(p_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \quad (4.8)$$

$$= \sum_{i=1}^I \sum_{j=1}^J \frac{1}{n} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} \quad (4.9)$$

$$= \sum_{i=1}^I \sum_{j=1}^J \frac{1}{n} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} \quad (4.10)$$

$$= \frac{\chi^2}{n} \quad (4.11)$$

Esercizio.

Verificare che le masse di colonna non sono altro che i profili medi di riga e che le masse di riga non sono altro che i profili medi di colonna. Calcolare le distanze al quadrato di ogni profilo riga e colonna dal profilo medio. Calcolare $in(I)$ e $in(J)$. Verificare che queste due quantità siano uguali. Costruire la matrice Z degli scostamenti standardizzati ponderati. Verificare che la somma dei quadrati degli elementi di Z è uguale a $in(I)$ e $in(J)$ ed è uguale alla somma dei quadrati dei valori singolari della matrice Z .

Soluzione

```
% Verifico che le masse di colonna non sono altro che i
% profili medi di riga e che le masse di riga non sono altro che i profili
% medi di colonna.
```

```

cchk=(ProfilesRows')*r;
rchk=ProfilesCols*c;
assert(max(abs(r-rchk))<1e-12,"Errore nel calcolo delle masse di riga")
assert(max(abs(c-cchk))<1e-12,"Errore nel calcolo delle masse di colonna")

% Distanza al quadrato dei profili riga dal profilo medio
% distI = vettore che contiene nell'elemento i-esimo la distanza (al
% quadrato) del profilo riga i-esimo dal profilo medio
distI=mahalFS(ProfilesRows,c',Dc);
% Distanza al quadrato dei profili colonna dal profilo medio
distJ=mahalFS(ProfilesCols',r',Dr);

% calcolo manuale di distI senza chiamare la funzione mahalFS
distIchk=zeros(I,1);
for i=1:I
    distIchk(i)=sum( (ProfilesRows(i,:)-c').^2./(c') );
end
assert(max(abs(distI-distIchk))<1e-12,"Errore di programmazione: le due inerzie non

% inI = inerzia totale dei punti riga
inI=sum(distI.*r);
% inJ = inerzia totale dei punti colonna

```

```

inJ=sum(distJ.*c);

assert(max(abs(distI-distIchk))<1e-12,"Errore di programmazione: le due " + .
    "inerzie non coincidono")

%% Costruzione la matrice Z (scostamenti standardizzati)

% Diversi modi di calcolo di Z

% zij = sqrt( (p_{ij} - r_ic_j)^2 / r_ic_j ) =(p_{ij} - r_ic_j)/sqrt(r_ic_j)
Z      = Dr^(-1/2) * (P - r * c') * Dc^(-1/2);
Zchk   = Dr^(1/2) * (ProfilesRows - onesI1 * c') * Dc^(-1/2);

assert(max(abs(Z-Zchk),[],'all')<1e-12,"Errore di programmazione nel calcolo

% Inerzia totale calcolata come somma dei quadrati degli elementi della
% matrice Z
intot=sum(Z.^2,'all');

%% SVD of Z
[U,Gam,V] = svd(Z,'econ');

% k = numero massimo di dimensioni latenti
k = min(I-1,J-1);

Gam = Gam(1:k,1:k);
U    = U(:,1:k);
V    = V(:,1:k);

```

```

%% Inerzia totale come somma dei quadrati valori singolari
% Gam2= matrice diagonale che contiene sulla diagonale principale i
% quadrati dei valori singolari
Gam2 = Gam.^2;
TotalInertia      = trace(Gam2);
assert(abs(TotalInertia-inI)<1e-12,"Errore di programmazione calcolo inerzia")

```

Il rango della matrice Z è al massimo pari al $\min(I-1, J-1)$ in questo caso

2. Questo significa che in questo esempio ci sono solo due dimensioni latenti e che i punti possono essere rappresentati nel piano cartesiano senza perdita di informazione.

La Tabella 4.9 riassume le caratteristiche dei punti riga e dei punti colonna, i rispettivi centroidi e l'inerzia totale nei due spazi.

4.2 Giudizi sulla bontà dell'analisi e punteggi

Nella sezione precedente abbiamo visto che la somma dei quadrati dei valori singolari coincide con l'inerzia totale. Di conseguenza, la somma dei quadrati dei primi due valori singolari diviso per la traccia di $Z'Z = \text{in}(I) = \text{in}(J)$, indica la quota di variabilità spiegata quando si rappresentano i punti nel piano formato dalle prime due componenti latenti. Similmente, il rapporto

$$\frac{\gamma_j^2}{\text{in}(I)}$$

Tabella 4.9: I profili riga e colonna definiscono due nuvole di punti nello spazio Euclideo ponderato di dimensione rispettivamente pari a $J - 1$ e $I - 1$

Punti	Nuvola dei punti riga I profili riga r_1, r_2, \dots, r_I nello spazio di dimensione $J - 1$.	Nuvola dei punti colonna J profili colonna c_1, c_2, \dots, c_J nello spazio di dimensione $I - 1$.
Masse	f_i frazione di unità statistiche nella riga i . $i = 1, 2, \dots, I$.	f_j frazione di unità statistiche nella colonna j . $j = 1, 2, \dots, J$.
Centroide	Il centroide dei profili riga è il vettore delle masse di colonna c .	Il centroide dei profili colonna è il vettore delle masse di riga r .
Metrica χ^2	Pesi dati dall'inversa degli elementi di c , (D_c^{-1}) .	Pesi dati dall'inversa degli elementi di r , (D_r^{-1}) .
Matrice profili standardizzati	$D_r^{1/2}(R - 1_{I \times 1}c')D_c^{-1/2} = D_r^{-1/2}(P - rc')D_c^{-1/2}$	$D_c^{1/2}(C - 1_{J \times 1}r')D_r^{-1/2} = D_c^{-1/2}(P' - cr')D_r^{-1/2}$
Inerzia totale dei punti	$in(I) = \sum_{i=1}^I f_i(r_i - c)'D_c^{-1}(r_i - c)$	$in(J) = \sum_{j=1}^J f_j(c_j - r)'D_r^{-1}(c_j - r)$
	$in(I) = in(J) = \sum_{s=1}^{\min(I-1, J-1)} \gamma_s^2 = \frac{\chi^2}{n}$	

indica il contributo delle generica j -esima componente latente all'inerzia totale. Come nell'analisi in componenti principali, si ottengono buoni risultati se con le prime due componenti latenti si riesce a spiegare una percentuale elevata dell'inerzia (pari almeno al 70%-75% del totale).

Dall'equazione (4.6) emerge che le coordinate dei punti riga sulla dimensione h esima sono date da

$$D_r^{-1/2} \gamma_h u_h$$

Similmente, le coordinate dei punti colonna sulla dimensione h esima sono date da

$$D_c^{-1/2} \gamma_h v_h$$

Queste coordinate vengono chiamate “coordinate principali”. Nella letteratura sull'analisi delle corrispondenze l'opzione più frequente è quello di rappresentare sia i punti riga sia i punti colonna in termini di coordinate principali.

Esercizio

Determinare la quota di varianza spiegata dalle diverse dimensioni latenti e rappresentare i punti riga e colonna in termini di coordinate principali. Calcolare le coordinate dei punti riga e colonna da inserire nel grafico che riporta le prime due dimensioni latenti. Rappresentare in un diagramma a dispersione gli score di riga e quelli di colonna utilizzando simboli diversi. Discutere i risultati ottenuti.

Soluzione

```
% cumsumTotalInertia = cumulative proportion of explained inertia
cumsumTotalInertia = cumsum(diag(Gam2))/TotalInertia;
```

```
% InertiaExplained è una matrice con quattro colonne.
% - La prima colonna contiene i valori singolari (la somma dei quadrati
%   dei valori singolari è l'inerzia totale = varianza totale della tabella
%   di contingenza)
% - La seconda colonna contiene gli autovalori (ossia i quadrati dei valori s
%   (la somma degli autovalori è l'inerzia totale)
% - La terza colonna contiene la varianza spiegata da ciascuna dimensione lat
% - La quarta colonna contiene la varianza cumulata spiegata da ciascuna dime
InertiaExplained=[diag(Gam) diag(Gam2) diag(Gam2 / TotalInertia) cumsumTotalI
ColNamesSummary={'Valori_singolari' 'Autovalori' 'Var_spiegata' 'Cum_Var_spie
RowNamesSummary=strcat(cellstr(repmat('dim_',k,1)), cellstr(num2str((1:k)'))))
RowNamesSummary=regexprep(RowNamesSummary,' ','');

InertiaExplainedtable=array2table(InertiaExplained,'VariableNames',ColNamesSu
    'RowNames',RowNamesSummary);
```

Il codice di cui sopra produce

	Valori_singolari	Autovalori	Var_spiegata	Cum_Var_spiegata
	-----	-----	-----	-----
dim_1	0.46972	0.22064	0.86191	0.86191
dim_2	0.18801	0.035349	0.13809	1

Le coordinate da inserire nel grafico si calcolano come segue.

```
% Coordinate principali dei punti riga
```

```
RowsPri      = Dr^(-1/2) * U*Gam;
```

```
% Coordinate principali dei punti colonna
```

```
ColsPri      = Dc^(-1/2) * V*Gam;
```

```
% Osservazione ColsPri'*Dc*ColsPri = matrice degli autovalori = quadrati
```

```
% dei valori singolari sulla diagonale principale
```

Il codice per produrre il grafico di analisi delle corrispondenze è riportato di seguito.

```
titl={'Grafico di analisi delle corrispondenze.' , ...
```

```
      'Plot of $X=D_r^{-1/2}U \backslash \Gamma$ and $Y= D_r^{-1/2} V \backslash \Gamma$'};
```

```
symbolrows='o';
```

```
symbolcols='^';
```

```
% Color for symbols and text for rows points
```

```
colorrows='b';
```

```
% Color for symbols and text for column rows
```

```
colorcols='r';
```

```
MarkerSize=14;
```

```
hold('on')
```

```
plot(RowsPri(:,1),RowsPri(:,2),'LineStyle','none','Marker',symbolrows,'Color', colorrows);
```

```
plot(ColsPri(:,1),ColsPri(:,2),'LineStyle','none','Marker',symbolcols,'Color', colorcols);
```

```
Lr=["A=Marca commerciale" "B=Marca industriale" "C=indifferente"];
```

```
Lc=["Liguria" "Lombardia" "Piemonte" "Veneto"];
```

```
dx=0.05;
```

```
text(RowsPri(:,1),RowsPri(:,2)+dx,Lr)
```

```

text(ColsPri(:,1),ColsPri(:,2)+dx,Lc)

FontName='Times';
FontSizeAxisLabels=12;
title(titl,'Interpreter','Latex');

% Inserisco nelle etichette degli assi la varianza spiegata
% L'istruzione sprintf('%5.1f,...') significa che il numero deve essere
% mostrato con una sola cifra decimale
xlabel(['Dimensione 1 (',sprintf('%5.1f',InertiaExplained(1,3)*100),'%')'], 'Fo
ylabel(['Dimension 2 (',sprintf('%5.1f',InertiaExplained(2,3)*100),'%')'], 'Fo
axis(gca,'equal')
xline(0);
yline(0)

```

L'output di questo codice è mostrato nella Figura 4.1.

La prossimità di due punti indicanti righe (colonne) indica un profilo simile nelle corrispondenti righe (colonne), dove “profilo” indica la distribuzione di frequenza condizionata delle riga (colonna); queste due righe (colonne) sono quindi quasi proporzionali. Interpretazione opposta si applica quando le due righe (colonne) sono invece distanti. La prossimità di un punto di riga con un punto di colonna indica che la riga (colonna) ha un peso particolarmente importante sulla colonna (riga). In contrapposizione a questo, un punto riga che si trova piuttosto distante da un particolare punto colonna indica che non ci sono quasi osservazioni nella colonna per quella riga (e viceversa). In altre parole, i punti riga che si trovano vicini a punti colonna rappresentano una

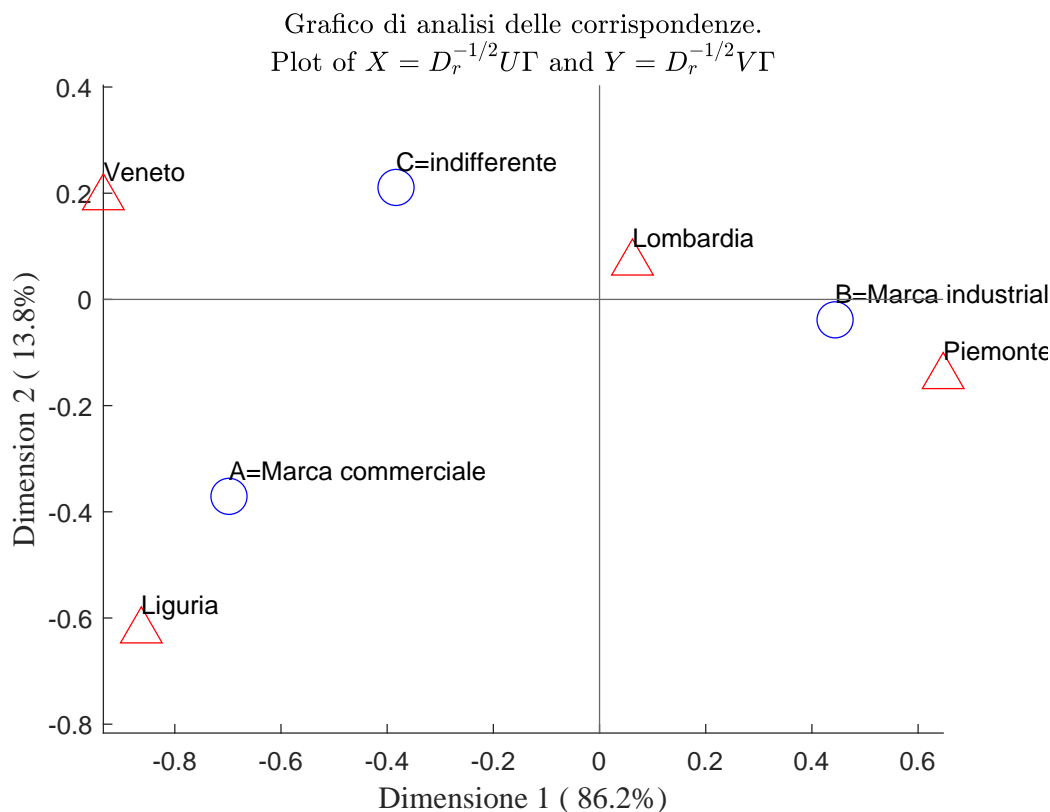


Figura 4.1: Grafico di analisi delle corrispondenze delle Tabella 4.5. I punti riga ed i punti colonna sono stati rappresentati utilizzando le coordinate principali.

combinazione riga/colonna che si presenta più di frequente di quanto atteso, qualora le variabili di riga e di colonna fossero indipendenti. Al contrario, punti riga e colonna che si trovano distanti tra loro indicano una cella nella tabella in cui la frequenza è inferiore rispetto a quanto ci si sarebbe atteso sotto l'ipotesi di indipendenza. Queste conclusioni poi sono particolarmente valide quando i punti si trovano distanti dall'origine degli assi. Le distanze euclidee nel grafico tra i punti riga oppure tra i punti colonna devono essere interpretate come distanze nelle metrica chi quadrato tra i profili riga oppure tra i profili colonna. L'origine degli assi è la media ponderata dei punti

riga e colonna. Di conseguenza, un punto (riga o colonna) proiettato che si trovi vicino all'origine indica un profilo vicino a quello medio. La Figura 4.1 mostra che la marca commerciale viene acquistata in prevalenza dai cittadini residenti in Liguria. Al contrario, il dentifricio di marca industriale viene utilizzato prevalentemente in Piemonte e Lombardia. Per i residenti in Veneto risulta per lo più indifferente quale tipologia di marca utilizzare. La prima dimensione latente può essere interpretata come la propensione ad utilizzare il dentifricio di marca industriale rispetto a quello di marca commerciale. La seconda componente latente, invece, può essere interpretata come il grado di indifferenza rispetto alla marca di dentifricio da utilizzare.

E' opportuno osservare che le prime due dimensioni latenti in questo caso spiegano il 100 per cento della variabilità (inerzia) complessiva.

4.3 Contributi all'inerzia del punto o all'inerzia della dimensione latente

Come le componenti principali avevano matrice di covarianza pari a $\Lambda = \Gamma\Gamma'$ la matrice di varianze e covarianze delle coordinate principali dei punti riga

4.3. CONTRIBUTI ALL'INERZIA DEL PUNTO O ALL'INERZIA DELLA DIMENSIONE LATENTE

e colonna è pari a $\Gamma\Gamma$.

$$\begin{aligned}
 cov(D_r^{-1/2}U\Gamma) &= (D_r^{-1/2}U\Gamma)'D_r(D_r^{-1/2}U\Gamma) \\
 &= \Gamma'U'U\Gamma \\
 &= \Gamma\Gamma = \Lambda \\
 cov(D_c^{-1/2}V\Gamma) &= (D_c^{-1/2}V\Gamma)'D_c(D_c^{-1/2}V\Gamma) \\
 &= \Gamma'V'V\Gamma \\
 &= \Gamma\Gamma = \Lambda
 \end{aligned}$$

La varianza ponderata del vettore di lunghezza I che contiene le coordinate dei punti riga per la componente h -esima (che indichiamo con $y_h^{(r)}$) è data da

$$\begin{aligned}
 y_h^{(r)'} D_r y_h^{(r)} &= (D_r^{-1/2}U_h\gamma_h)'D_r(D_r^{-1/2}U_h\gamma_h) = \lambda_h \\
 &= \sum_{i=1}^r y_{ih}^{(r)} f_i.
 \end{aligned}$$

Di conseguenza, il contributo della riga i -esima all'asse h è dato da:

$$\frac{\left(y_{ih}^{(r)}\right)^2 f_i}{\lambda_h}$$

In maniera analoga, il contributo della colonna j -esima all'asse h è dato da:

$$\frac{\left(y_{jh}^{(c)}\right)^2 f_{.j}}{\lambda_h}$$

Similmente, ci possiamo chiedere in quale misura l'inerzia del punto (ossia la sua distanza rispetto al profilo medio nella metrica chi quadrato) viene

spiegata dalle diverse dimensioni latenti. La somma dei quadrati degli elementi di ogni riga nella matrice degli scores principali dei punti riga è uguale alla rispettiva distanza al quadrato di ogni profilo riga dal centroide secondo la metrica chi quadrato. In simboli matriciali

$$\begin{aligned} \text{diag}(D_r^{-1/2} U \Gamma^2 U' D_r^{-1/2}) &= \text{diag}(\chi^2 d_{1c}^2, \dots, \chi^2 d_{Ic}^2) \\ &= \text{diag}((r_1 - c)' D_r^{-1} (r_1 - c), \dots, (r_I - c)' D_r^{-1} (r_I - c)) \\ (y_{i1}^{(r)})^2 + \dots + (y_{ik}^{(r)})^2 &= \chi^2 d_{ic}^2 \end{aligned}$$

Il rapporto

$$\frac{(y_{ih}^{(r)})^2}{\chi^2 d_{ic}^2} = \frac{(y_{ih}^{(r)})^2 f_i}{\chi^2 d_{ic}^2 f_i}$$

fornisce il contributo della dimensione h -esima alla spiegazione dell'inerzia del punto riga i . Similmente il rapporto

$$\frac{(y_{jh}^{(c)})^2}{\chi^2 d_{jr}^2} = \frac{(y_{jh}^{(c)})^2 f_j}{\chi^2 d_{jr}^2 f_j}$$

fornisce il contributo della dimensione h -esima alla spiegazione dell'inerzia del punto colonna j .

Esercizio: calcolare il contributo dei punti riga e colonna alla spiegazione delle due dimensioni latenti. Calcolare il contributo delle due diverse dimensioni alla spiegazione dell'inerzia di ogni punto riga e colonna.

```
%% Analisi dei contributi
```

```
% Osservazione RowsPri'*Dr*RowsPri = matrice degli autovalori = quadrati
```

```
% dei valori singolari sulla diagonale principale
```


4.3. CONTRIBUTI ALL'INERZIA DEL PUNTO O ALL'INERZIA DELLA DIMENSIONE LATENTE

```
disp("Contributi delle diverse righe alla spiegazione del primo autov.")
```

```
disp(RowsPri(:,1).^2.*r/(Gam(1,1)^2))
```

```
% Osservazione: la quantità
```

```
% sum(RowsPri(:,1).^2.*r)
```

```
% è la varianza ponderata della prima dimensione (primo autovalore =Gam(1,1)^2)
```

```
disp("Contributi delle diverse righe alla spiegazione del secondo autov.")
```

```
disp(RowsPri(:,2).^2.*r/(Gam(2,2)^2))
```

```
% Osservazione: la quantità
```

```
% sum(RowsPri(:,2).^2.*r)
```

```
% è la varianza ponderata della seconda dimensione (secondo autovalore =Gam(2,2)^2)
```

```
% Contributi delle diverse colonne alla determinazione del primo autovalore
```

```
disp("Contributi delle diverse colonne alla spiegazione del secondo autov.")
```

```
disp(ColsPri(:,1).^2.*c/(Gam(1,1)^2))
```

```
% Osservazione: la quantità
```

```
% sum(ColsPri(:,1).^2.*c)
```

```
% è la varianza ponderata della prima dimensione (primo autovalore =Gam(1,1)^2)
```

```
% Contributi delle diverse colonne alla determinazione del secondo autovalore
```

```
disp("Contributi delle diverse righe alla spiegazione del secondo autov.")
```

```
disp(ColsPri(:,2).^2.*c/(Gam(2,2)^2))
```

```
% Osservazione: la quantità
```

```

% sum(ColsPri(:,2).^2.*c)
% è la varianza ponderata della seconda dimensione (secondo autovalore =Gam(2)

%% Calcolare i contributi della dimensione all'inerzia dei punti
% Osservazione: il vettore distI.^2  contiene la distanza al quadrato di
% ogni profilo riga dal profilo medio (inerzia di ogni punto)
% Il vettore
% Contributi della dimensione 1 alla spiegazione dell'inerzia di ogni punto
disp("Contributi delle due dimensioni alla spiegazione dell'inerzia..." + ...
      " di ogni punto riga")
disp(RowsPri(:,1:2).^2./distI.^2)
disp("Contributi delle due dimensione alla spiegazione dell'inerzia..." + ...
      " di ogni punto colonna")
disp(ColsPri(:,1:2).^2./distJ.^2)

```

Il quadro completo della scomposizione dell'inerzia totale è riportato nella Tabella 4.10.

4.3. CONTRIBUTI ALL'INERZIA DEL PUNTO O ALL'INERZIA DELLA DIMENSIONE LATENTE

		Asse 1	Asse 2	...	Asse k	Inerzia del punto
righe	1	$f_{1.} \left(y_{11}^{(r)}\right)^2$	$f_{1.} \left(y_{12}^{(r)}\right)^2$...	$f_{1.} \left(y_{1k}^{(r)}\right)^2$	$f_{1.} \sum_{j=1}^k \left(y_{1j}^{(r)}\right)^2$
	2	$f_{2.} \left(y_{21}^{(r)}\right)^2$	$f_{2.} \left(y_{22}^{(r)}\right)^2$...	$f_{2.} \left(y_{2k}^{(r)}\right)^2$	$f_{2.} \sum_{j=1}^k \left(y_{2j}^{(r)}\right)^2$
	\vdots	\vdots	\vdots	...	\vdots	\vdots
	I	$f_{I.} \left(y_{I1}^{(r)}\right)^2$	$f_{I.} \left(y_{I2}^{(r)}\right)^2$...	$f_{I.} \left(y_{Ik}^{(r)}\right)^2$	$f_{I.} \sum_{j=1}^k \left(y_{Ij}^{(r)}\right)^2$
	In dim	$\sum_{i=1}^k \left(y_{i1}^{(r)}\right)^2 f_{i.}$ $= \lambda_1 = \gamma_1^2$	$\sum_{i=1}^k \left(y_{i2}^{(r)}\right)^2 f_{i.}$ $= \lambda_2 = \gamma_2^2$...	$\sum_{i=1}^k \left(y_{iJ}^{(r)}\right)^2 f_{i.}$ $= \lambda_k = \gamma_k^2$	$\sum_{i=1}^k \lambda_j =$ $in(I) = in(J)$
colonne	1	$f_{.1} \left(y_{11}^{(c)}\right)^2$	$f_{.1} \left(y_{12}^{(c)}\right)^2$...	$f_{.1} \left(y_{1k}^{(c)}\right)^2$	$f_{.1} \sum_{j=1}^k \left(y_{1j}^{(c)}\right)^2$
	2	$f_{.2} \left(y_{21}^{(c)}\right)^2$	$f_{.2} \left(y_{22}^{(c)}\right)^2$...	$f_{.2} \left(y_{2k}^{(c)}\right)^2$	$f_{.2} \sum_{j=1}^k \left(y_{2j}^{(c)}\right)^2$
	\vdots	\vdots	\vdots	...	\vdots	\vdots
	J	$f_{.J} \left(y_{J1}^{(c)}\right)^2$	$f_{.J} \left(y_{J2}^{(c)}\right)^2$...	$f_{.J} \left(y_{Jk}^{(c)}\right)^2$	$f_{.J} \sum_{j=1}^k \left(y_{Jj}^{(c)}\right)^2$

Tabella 4.10: Scomposizione dell'inerzia totale nelle k dimensioni latenti. La somma di ogni riga è l'inerzia del punto. Ogni entrata della tabella diviso il totale dell'ultima colonna è il contributo della dimensione all'inerzia del punto. La somma di ogni colonna (distinta per punti righe e colonna) è l'inerzia della dimensione. Ogni entrata della tabella divisa per il corrispondente totale nella riga denominata "In dim" (In dim = inerzia della dimensione) è il contributo del punto all'inerzia della dimensione.