Capitolo 5

Analisi dei grappoli

5.1 Introduzione

Sotto il nome analisi dei grappoli (*Cluster Analysis*) vengono raccolte un insieme di tecniche statistiche che presentano il comune obiettivo di effettuare raggruppamenti di unità statistiche in base alla similarità del loro profilo, descritto da un insieme di variabili. I gruppi risultanti dovrebbero essere caratterizzati da un elevato grado di omogeneità interna e vi dovrebbe essere una altrettanto elevata disomogeneità tra i gruppi. La ragione per cui si realizza tale operazione può variare: intento classificatorio, riduzione della dimensionalità dei dati, analisi esplorativa, etc. Nel marketing la tecnica trova largo impiego per la segmentazione a posteriori del mercato.

Un punto preliminare, già discusso nella sezione 4.2 riguarda la selezione delle variabili che definiscono il profilo individuale delle unità: l'inclusione di variabili poco discriminanti o poco rilevanti al fine della caratterizzazione dei gruppi esercita un ruolo negativo sulla qualità dei risultati.

Effettuate le misurazione sugli individui, va affrontato il problema di scegliere una misura di (dis)-similarità, secondo le linee discusse nel capitolo precedente. Un problema particolare sorge quando le variabili presentano una scala molto diversa (campo di variazione ed unità di misura). In generale, l'importanza relativa di ciascuna variabile nella formazione dei grappoli è direttamente collegata alla varianza delle diverse variabili. Pertanto, variabili caratterizzate da un più elevato grado di dispersione hanno maggiore impatto sulla misura di distanza (es: attitudine verso un prodotto, età e reddito in lire. Si andrebbero ad individuare gruppi prevalentemente sulla base delle differenze di reddito).

Un possibile rimedio consiste nella standardizzazione delle variabili, mediante sottrazione della media e divisione per la deviazione standard, o l'impiego di una misura di distanza ponderata o normalizzata. Un caso particolare è la distanza di Mahalanobis, che consente di eliminare l'effetto dovuto alla presenza di variabili correlate sulla misura della dissimilarità.

Le $n(n-1)/2$ distanze vengono raccolte nella matrice simmetrica:

$$\boldsymbol{D} = \begin{bmatrix} 0 & d_{12} & \cdots & \cdots & d_{1n} \\ & 0 & & & d_{2n} \\ & & \ddots & & \vdots \\ & & & \ddots & d_{n-1,n} \\ & & & & 0 \end{bmatrix} \tag{5.1}$$

5.2 Metodi di raggruppamento delle unità

I metodi di raggruppamento si distinguono in gerarchici e partitivi (o non gerarchici); dal momento che soltanto i primi seguono una sequenza ordinata di operazioni della stessa natura. I secondi richiedono che il numero dei gruppi sia determinato a priori e forniscono un'unica partizione come risultato finale. I metodi gerarchici sono ulteriormente distinti in

1. *agglomerativi*: procedono per aggregazioni successive delle unità partendo da n gruppi formati da un solo individuo.
2. *divisivi*: partono da un solo gruppo formato da tutte le unità e procedono a partizioni successive fino a giungere a gruppi formati da una sola unità.

5.3 Metodi gerarchici agglomerativi

La struttura logica dei metodi agglomerativi può essere così sintetizzata:

1. Nello stadio iniziale ciascuna unità costituisce un gruppo separato. La distanza tra i gruppi è fornita dalla matrice $\boldsymbol{D}$.
2. I due gruppi che possiedono distanza minima vengono fusi; la distanza a cui avviene la fusione viene registrata.
3. Si calcola la distanza tra il nuovo gruppo, sorto dalla fusione di cui al punto precedente, e i gruppi gi à esistenti. Si eliminano 2 righe e colonne dalla matrice $\boldsymbol{D}$ in corrispondenza dei gruppi fusi e vengono rimpiazzate da una singola riga e colonna che contengono le nuove distanze. La dimensione della matrice $\boldsymbol{D}$ si riduce di una unità.
4. Vengono ripetuti i passi 2 e 3 finchè non si giunge ad una configurazione in cui esiste un solo gruppo (ciò richiede $(n-1)$ iterazioni). Il processo di fusione rappresentato graficamente attraverso il dendrogramma: questo riporta sull'asse orizzontale il livello di distanza a cui avviene la fusione e sull'asse delle ascisse riporta le unità. Ad ogni livello di distanza corrisponde una partizione.

I metodi proposti differiscono per le modalità di calcolo della distanza tra gruppi al punto 3. Lo strumento grafico che consente di sintetizzare il processo di fusione è il dendrogramma, dal quale è anche possibile apprezzare quanto un gruppo sia separato dagli altri. Il rapporto tra il livello di distanza a cui un gruppo viene formato e quello a cui si fonde con un altro può essere utilizzato al fine di individuare il numero dei grappoli, poiché è tanto più elevato quanto più il grappolo è delimitato e separato dai rimanenti.

5.3.1 Il metodo del legame singolo (*nearest neighbour*)

La distanza tra gruppi è misurata dalla distanza più piccola esistente tra gli elementi appartenenti ad un gruppo e quelli appartenenti ad un altro.

A titolo illustrativo consideriamo 5 oggetti A, B, C, D, E, la cui matrice di distanze è:

$$\begin{array}{c} \\ (A) \\ (B) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{ccccc} (A) & (B) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & & \\ \boxed{2} & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array} \tag{5.2}$$

la coppia di unità che presenta distanza minima è (AB); le medesime sono unite in un gruppo. Si deve ora determinare la distanza tra il gruppo appena formato e le rimanenti unità; questa sarà fornita dalla più piccola delle distanze con le unità compoenenti il gruppo (ad esempio, la distanza tra (AB) e (C) è uguale al minore tra 6 e 5).

$$\begin{array}{c} \\ (AB) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 9 & 4 & 0 & \\ 8 & 5 & \boxed{3} & 0 \end{pmatrix} \end{array}$$

A questo punto le unità (D) e (E) vengono fuse nel gruppo (DE) e si ottiene la nuova matrice di distanze:

$$\begin{array}{c} \\ (AB) \\ (C) \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & (C) & (DE) \end{array} \\ \begin{pmatrix} 0 & & \\ 5 & 0 & \\ 8 & \boxed{4} & 0 \end{pmatrix} \end{array}$$

Vengono fusi i gruppi (C) e (DE), che presentano distanza minima (4), ottenendosi

$$\begin{array}{c} \\ (AB) \\ (CDE) \end{array} \begin{array}{c} \begin{array}{cc} (AB) & (CDE) \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{5} & 0 \end{pmatrix} \end{array}$$

L'ultima iterazione aggrega i due gruppi in un unico gruppo contenente tutte le unità. La sequenza delle fusioni è pertanto rappresentata nella tabella seguente:

Iterazione	Gruppi	Livello di distanza
0	$(A)(B)(C)(D)(E)$	-
1	$(AB)(C)(D)(E)$	2
2	$(AB)(C)(DE)$	3
3	$(AB)(CDE)$	4
4	$(ABCDE)$	5

Il dendrogramma corrispondente è presentato nella figura 5.1, nel riquadro in alto a sinistra.

Una caratteristica (ed anche un limite) del metodo sta nel produrre tendenzialmente dei grappoli allungati (a salciccia) in relazione al fatto che la fusione dei gruppi avviene facendo riferimento ad un solo legame. Quando esistono grappoli ben delineati, ma non separati, il concatenamento potrebbe indurre a considerare un unico grappolo. Tuttavia il metodo consente di individuare grappoli di qualsiasi forma e mette in luce eventuali valori anomali meglio di altre tecniche.

5.3.2 Metodo del legame completo (*furthest neighbour*)

In base a questo metodo la distanza tra i gruppi è definita come la massima distanza esistente tra gli individui componenti. Questa rappresenta il diametro della sfera che contiene tutti i punti appartenenti ai due gruppi. Con riferimento all'esempio precedente, il primo passo, basato sulla matrice originaria delle distanze (5.2), è identico e porta alla formazione del grappolo (AB). Le differenze sorgono ora con riferimento al calcolo della distanza tra (AB) e le altre unità. Ad esempio, quella tra (AB) e (C) sarà fornita dal più grande tra i valori $d_{AC} = 6$ e $d_{BC} = 5$.

$$\begin{array}{c} \\ (AB) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & \\ 6 & 0 & & \\ 10 & 4 & 0 & \\ 9 & 5 & \boxed{3} & 0 \end{pmatrix} \end{array}$$

Nuovamente, le unità (D) e (E) vengono fuse nel gruppo (DE) e si perviene a:

$$\begin{array}{c} \\ (AB) \\ (C) \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & (C) & (DE) \end{array} \\ \begin{pmatrix} 0 & & \\ 6 & 0 & \\ 10 & \boxed{5} & 0 \end{pmatrix} \end{array}$$

Si fondono ora i gruppi (C) e (DE), che presentano distanza minima (5),

$$\begin{array}{c} \\ (AB) \\ (CDE) \end{array} \begin{array}{c} \begin{array}{cc} (AB) & (CDE) \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{10} & 0 \end{pmatrix} \end{array}$$

L'ultima iterazione aggrega i due gruppi in un unico gruppo contenente tutte le unità. Si noti che i cluster non cambiano rispetto al caso precendente, ma variano i livelli di distanza a cui vengono effettuate le aggregazioni; in particolare, risulta più accentuato il salto nel livello di distanza al quale avviene l'ultima fusione. Il dendrogramma corrispondente è presentato nella figura 5.1, nel riquadro in alto a destra.

5.3.3 Metodo del legame medio (*average linkage*)

La distanza tra gruppi è calcolata come media aritmetica semplice delle distanze tra tutte le unità che compongono i due gruppi. Con riferimento a (5.2) la distanza tra il gruppo (AB) e (C) è la media aritmetica semplice tra i valori $d_{AC} = 6$ e $d_{BC} = 5$, e pertanto alla prima iterazione:

$$\begin{array}{c} \\ (AB) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & \\ 5.5 & 0 & & \\ 9.5 & 4 & 0 & \\ 8.5 & 5 & \boxed{3} & 0 \end{pmatrix} \end{array}$$

Le iterazioni successive forniscono:

$$\begin{array}{c} \\ (AB) \\ (C) \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & (C) & (DE) \end{array} \\ \begin{pmatrix} 0 & & \\ 5.5 & 0 & \\ 9 & \boxed{4.5} & 0 \end{pmatrix} \end{array}$$

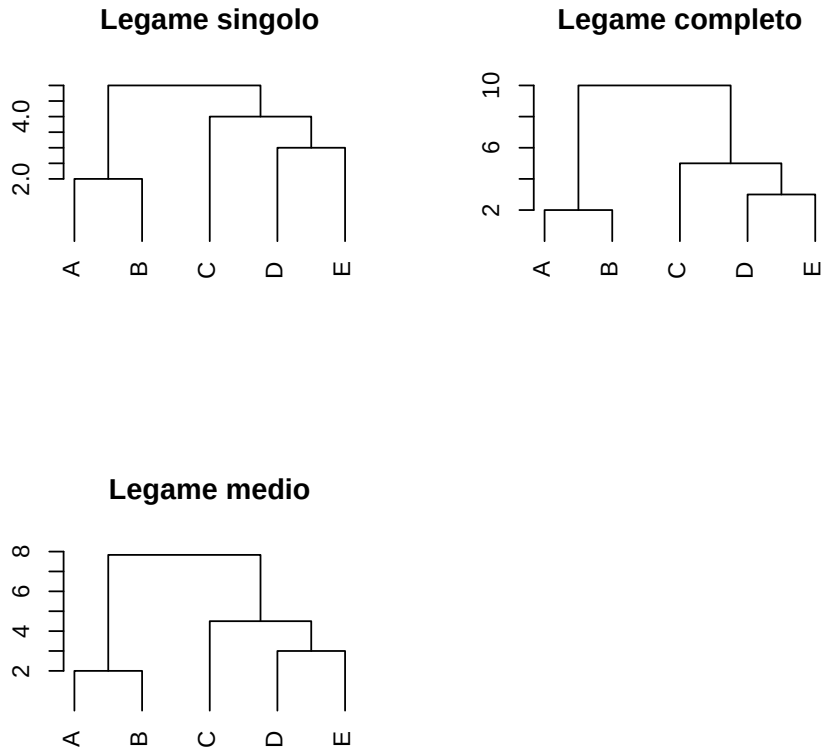
$$\begin{array}{c} \\ (AB) \\ (CDE) \end{array} \begin{array}{c} \begin{array}{cc} (AB) & (CDE) \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{7.25} & 0 \end{pmatrix} \end{array}$$

Si osservi che la fusione dei gruppi avviene a livelli di distanza intermedi tra quelli che caratterizzano i due metodi precedenti. Il dendrogramma corrispondente è presentato nella figura 5.1, nel riquadro in basso a sinistra.

5.3.4 Metodo del centroide

La distanza tra i gruppi è pari alla distanza tra i centroidi, vale a dire i valori medi calcolati sugli individui appartenenti ai gruppi. Tale metodo richiede quindi la matrice dei dati

Figura 5.1: Analisi dei grappoli: dendrogramma per quattro metodi gerarchici



originaria, $\boldsymbol{X}$. Dà luogo a fenomeni gravitazionali, per cui i gruppi grandi tendono ad attrarre al loro interno i gruppi piccoli. Inoltre le distanze a cui avviene le successive fusioni possono essere non crescenti.

5.3.5 Metodo di Ward

Si fonda sulla scomposizione della devianza totale in devianza tra i grappoli e varianza entro i grappoli. Ad ogni passo l'unione di tutte le possibili coppie di *cluster* è considerata e viene fusa la coppia per cui la varianza entro i gruppi è minima. Tende a produrre cluster che hanno pressappoco lo stesso numero (limitato) di osservazioni.

5.3.6 L'analisi gerarchica in R

L'analisi gerarchica viene effettuata mediante la funzione

```
hclust(d, method = "complete")
```

che ha come input la matrice di distanze; i metodi disponibili sono quello del legame singolo (`single`), del legame completo (`complete`), del legame medio (`average`), e altri metodi.

Output della funzione hclust:

- `$merge`: sequenza del processo di fusione,
- `$height`: vettore che indica il livello di distanza attraverso il quale è avvenuta la fusione, la cui lunghezza equivale al numero di iterazioni,
- `$order`: opportuna permutazione delle unità finalizzata alla costruzione del dendrogramma.
- `$labels`: l'etichetta che contrassegna le unità

Il dendrogramma è fornito dalla funzione

```
plot.hclust(hclust.obj, labels, hang = 0.1, ...)
```

Al fine di scegliere la partizione del dendrogramma, si può utilizzare il vettore `$height` generato dalla funzione hclust, calcolando le grandezze

$$\frac{d_k}{d_{k-1}}, k = 1, 2, ..., m$$

dove d_k rappresenta il livello di distanza a cui è stata effettuata la fusione al passo k e m il numero di iterazioni effettuate. Il rapporto risulta utile nella scelta del numero dei grappoli. Quando esso risulta sufficientemente elevato, significa che i gruppi sono sufficientemente dissimili tra di loro, per cui è possibile tagliare il dendrogramma a livello di distanza corrispondente.

Presentiamo ora una applicazione con riferimento al data set `mtcars`, considerato nel capitolo precedente e contenente 13 misurazioni di diversi aspetti tecnici e attinenti la performance riferite a 32 autoveicoli (maggiori dettagli possono essere ottenuti invocando `help(mtcars)`.

```
>library(mva)
>data(mtcars)
>help(mtcars)
>x <- scale(mtcars[,1:7])
>d <- dist(x)
```

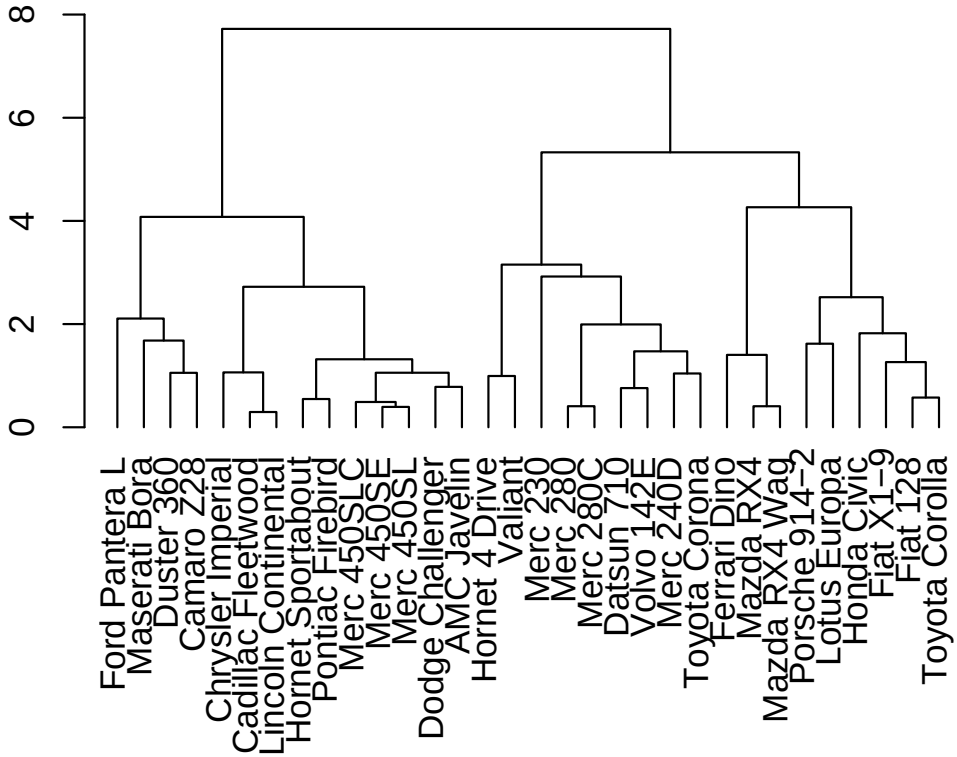
```
>lc <- hclust(d,method="complete")
>lc
$merge
      [,1] [,2]
 [1,]  -15  -16
 [2,]  -12  -13
 [3,]   -1   -2
 [4,]  -10  -11
 ...    ..   ..
 ...    ..   ..
 ...    ..   ..
[30,]   27   29
[31,]   28   30

$height
 [1] 0.2956825 0.3944266 0.4075899 0.4082884 0.4901305 0.5475333 0.5757917
 [8] 0.7595603 0.7827694 0.9936969 1.0428738 1.0554323 1.0566522 1.0635310
[15] 1.2631917 1.3181107 1.4032977 1.4721123 1.6199219 1.6809662 1.8220229
[22] 1.9934625 2.1075394 2.5210420 2.7226786 2.9221444 3.1529877 4.0778628
[29] 4.2649123 5.3291587 7.7221893

$order
 [1] 29 31  7 24 17 15 16  5 25 14 12 13 22 23 4 6 9 10 11 3 32 8 21 30 1
[26]  2 27 28 19 26 18 20

>ls <- hclust(d,method="single")
>plot.hclust(lc,-1)
```

I rapporti d_k/d_{k-1}, $k = 1, 2, \ldots, n-1$, segnalano un salto in corrispondenza dell'ultima aggregazione ($d_{n-1}/d_{n-2} = 1.45$); questa evidenza potrebbe essere presa a supporto della individuazione di 2 grappoli di unità. Per esercizio si confronti la soluzione del legame completo con il legame singolo, il quale non supporta una ripartizione in due gruppi.

5.4 Metodi gerarchici divisivi

Nei metodi gerarchici divisivi la configurazione iniziale prevede che tutte le unità siano raggruppate in un unico cluster. Al primo passo l'insieme di n unità viene suddiviso in due gruppi: dal momento che esistono $2^{n-1} - 1$ possibili soluzioni, si pone un problema computazionale ed occorre imporre delle restrizioni per avere una soluzione praticabile. Nei metodi cosiddetti nodali, si procede alla scelta delle due unità più distanti come nodi o fuochi e tutte le altre unità vengono allocate ai due gruppi in ragione della loro vicinanza

Figura 5.2: Analisi dei grappoli: metodo del legame completo per il data set `mtcars`



rispetto ai nodi. Successivamente vengono suddivisi i due grappoli con lo stesso criterio e si continua il processo finché ciascuna unità costituisce un gruppo a parte. L'algoritmo, che consiste di $n-1$ divisioni successive, può essere così descritto:

1. si individua una coppia di punti nodali, (punti che presentano distanza massima);
2. si attribuiscono le unità rimanenti ai due gruppi corrispondenti ai punti nodali, in base alla distanza minima dai punti nodali;
3. si iterano i passi precedenti (all'interno dei nuovi gruppi si individuano due punti nodali, etc.) finché si avranno n gruppi.

Con riferimento all'esempio numerico precedente, si ha che i punti che distano maggiormente sono A e D ($d_{AD} = 10$). Pertanto si otterrà la prima partizione $[(AB), (CDE)]$. Le matrici di distanze tra gli elementi dei due gruppi sono

$$\begin{array}{c} \\ A \\ B \end{array}\!\!\begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{2} & \end{pmatrix} \end{array}, \qquad \begin{array}{c} \\ C \\ D \\ E \end{array}\!\!\begin{array}{c} \begin{array}{ccc} C & D & E \end{array} \\ \begin{pmatrix} 0 & & \\ 4 & 0 & \\ \boxed{5} & 3 & 0 \end{pmatrix} \end{array}$$

Alla seconda iterazione, le unità A e B vanno a costituire due gruppi separati e C ed E vengono scelti come punti nodali. Si ottiene la seconda partizione: A,B,(CD),E. Alla terza ed ultima iterazione, ciascuna unità costituisce un gruppo a se stante.

5.5 Metodi non gerarchici

Richiedono che il numero dei cluster sia specificato a priori e generalmente forniscono una sola partizione come output. Il metodo più popolare, dovuto a Hartigan e Wong [11], prende il nome di k-means e consiste nello specificare k punti iniziali, o *seeds* (scegliendo in maniera opportuna alcune unità o prendendo la configurazione determinata da una tecnica gerarchica). Al primo passo ciascuna unità è assegnata ai k punti in ragione della distanza più piccola. Viene calcolata la media o il centroide per ciascuno dei k gruppi e si verifica che ciascuna unità sia assegnata al grappolo che ha il centroide più vicino. Se questo non si verifica si procede a spostare l'unità presso il grappolo che ha il centroide più vicino e si controlla la nuova soluzione, per cui si procede iterativamente a spostamenti successivi fino a raggiungere una configurazione stabile.

Alternativamente si puo utilizzare per la riallocazione delle unità il criterio di minimizzare la varianza entro i gruppi. La configurazione finale e la velocità di convergenza dell'algoritmo dipendono dalla scelta dei seeds, per cui, se non si hanno informazioni a priori sufficientemente affidabili, è consigliabile applicare l'algoritmo con valori iniziali diversi, e controllare la stabilità della soluzione ottenuta. Questa si rivelerà molto instabile quando la popolazione di riferimento è omogenea e non ammette partizioni.

5.5.1 Il metodo PAM

L'algoritmo PAM (*Partitioning around Medoids*), proposto da [10] si fonda sulla ricerca di k punti rappresentativi, detti medoidi, tra quelli osservati; le restanti unità sono allocate ai medoidi in ragione della distanza più piccola. La media delle distanze dal medoide più vicino misura la bontà della soluzione ottenuta. L'obiettivo finale è quello di giungere ad una partizione che minimizza la somma delle distanze entro i gruppi.

La fase di identificazione preliminare dei medoidi (detta BUILD) parte dalla selezione dell'unità per la quale risulta minima la somma delle distanze o dissimilarità da tutte le altre unità. I rimanenti $k-1$ punti sono individuati mediante la seguente procedura: per $j = 2, \ldots, k$

1. si denoti con S_j il set dei medoidi selezionati al passo j; per ogni unità $i \notin S_j$:
2. per ogni punto $l \neq i$ e $\notin S_j$ si calcolano la distanza più piccola dai punti in S_j, che denotiamo D_l, la distanza da i, d_{il} e $C_{il} = \max\{D_l - d_{il}, 0\}$.
3. Si seleziona l'unità che massimizza la somma

$$\sum_{l \notin S_j, l \neq i} C_{il}$$

La procedura mira a garantire che il medoide candidato disti il più possibile da quelli già scelti. Contribuiscono alla funzione criterio i punti vicini a quello candidato, come implica l'operazione di massimo che definisce C_{il}.

La seconda fase, detta *SWAP*, mira a migliorare la configurazione preliminare. Si considerano le $k \times (n-k)$ coppie $(i,h), i \in S_k, h \notin S_k$, formate da un medoide selezionato nella fase build e tutte le restanti; si cercano di valutare le conseguenze di un rovesciamento di ruoli, uno *swap*, tra le unità in S_k e quelle esterne.

Per ogni unità j diversa da quelle candidate per lo swap, i e h, si denoti con D_j la distanza dal punto in S_k più vicino, $D_j = \min_{l \in S_k}(d_{lj})$

si definisce il contributo allo scambio dell'unità i con h, $C_j^{(ih)}$, nella maniera seguente:

i. se $\min(d_{ij}, d_{hj}) > \min_{r \in S_k, r \neq i}(d_{rj})$, $C_j^{(ih)} = 0$ (il contributo è nullo se j è sufficientemente remoto dai candidati; dal punto di vista di j è indifferente lo scambio dei candidati, dal momento che fa riferimento ad un altro punto rappresentativo);

ii. se $d_{ij} = D_j = \min_{r \in S_k}(d_{rj})$ (il punto j già appartiene al cluster del candidato i), possono sussistere due casi:

- **ii.1.** se inoltre $d_{hj} < \min_{r \in S_k, r \neq i}(d_{rj})$, $C_j^{(ih)} = d_{hj} - d_{ij}$ (si osservi che se j si trova tra h e j il contributo risulta negativo),
- **ii.2.** altrimenti $C_j^{(ih)} = \min_{r \in S_k, r \neq i}(d_{rj}) - D_j$ (il contributo risulta sempre positivo, sfavorevole allo scambio di i con j, poiché $D_j < \min_{r \in S_k, r \neq i}(d_{rj})$)

iii. se $d_{ij} > D_j = \min_{r \in S_k}(d_{rj})$, $C_j^{(ih)} = d_{hj} - D_j$ (in questo caso il punto j contribuisce negativamente - è favorevole - allo scambio di i con h)

Si denoti ora con $T^{(ih)} = \sum_j C_j^{(ih)}$, il contributo di tutti i punti diversi da i e h allo scambio tra i due candidati; si seleziona la coppia (i, h) per quale $T^{(ih)}$ risulta minimo. Se $T^{(ih)} < 0$ si effettua lo scambio e la procedura ripetuta fino a quando $T^{(ih)} \geq 0$

5.5.2 Esempio

L'analisi dei grappoli non gerarchica con metodo k-means in R si effettua mediante la funzione

```
kmeans(x, centers, iter.max=10)
```

dove i valori iniziali (`centers`) possono essere derivati preliminarmente attraverso una tecnica gerarchica, ovvero possono essere determinati vengono determinati casualmente dal programma, nel qual caso `centers` è posto pari al numero desiderato di gruppi.

L'output della funzione comprende:

- `km$cluster`: vettore di allocazione delle unit à,
- `km$center`: matrice dei centroidi,
- `km$withinss`: varianze entro i gruppi
- `km$size`: dimensione dei gruppi.

Ad esempio, con riferimento al data set `mtcars` si utilizza la partizione ottenuta dall'analisi gerarchica con il metodo del legame completo, con l'individuazione di tre gruppi.

```
>initial <- tapply(x,list(rep(cutree(lc,3),ncol(x)),col(x)),mean)
>km <- kmeans(x,initial,100)
>km

$cluster
 [1] 2 2 2 2 3 2 3 2 2 2 2 3 3 3 3 3 1 1 1 2 3 3 3 3 1 1 1 3 2 3 2

$centers
         mpg        cyl       disp         hp       drat         wt       qsec
1  1.6552394 -1.2248578 -1.1624447 -1.0382807  1.2252295 -1.3738462  0.3075550
2  0.1384407 -0.5716003 -0.5707543 -0.5448163  0.1887816 -0.2454544  0.5491221
3 -0.8280518  1.0148821  0.9874085  0.9119628 -0.6869112  0.7991807 -0.6024854

$withinss
[1]  7.76019 28.61309 33.37849

$size
[1]  6 12 14
```

La funzione `cutree` taglia il dendrogramma in relazione al numero dei gruppi indicato in argomento e fornisce tutte le informazioni necessarie per allocare le unità ai gruppi. Di

Va osservato, comunque, che l'analisi gerarchica, effettuata con il metodo del legame completo, non supportava la divisione in tre gruppi; in effetti, se ripetiamo l'applicazione partendo da diversi punti iniziali, utilizzando, ad esempio,

```
>km <- kmeans(x,3,100)
```

l'algoritmo k-means converge ad una soluzione diversa; questo potrebbe essere interpretato come il riflesso dell'assenza di una partizione naturale in tre gruppi. Ripetendo l'esercizio specificando soltanto due gruppi, si ha l'interessante risultato che, indipendentemente dalla scelta dei punti iniziali, l'algoritmo converge alla soluzione:

```
>  kmeans(x,2,100)
$cluster
 [1] 2 2 2 2 1 2 1 2 2 2 2 1 1 1 1 1 2 2 2 2 1 1 1 2 2 2 1 2 1 2

$centers
         mpg        cyl       disp         hp       drat         wt       qsec
1 -0.8280518  1.0148821  0.9874085  0.9119628 -0.6869112  0.7991807 -0.6024854
2  0.6440403 -0.7893528 -0.7679844 -0.7093044  0.5342642 -0.6215850  0.4685997

$withinss
[1] 33.37849 59.28078

$size
[1] 14 18
```

Lasciando al lettore il confronto con la soluzione gerarchica, si rileva che l'interpretazione dei risultati e la caratterizzazione dei grappoli va effettuata guardando ai centroidi dei due gruppi, i quali possono evidenziare i diversi profili dei gruppi; ad esempio, il secondo gruppo contiene gli autoveicoli con minore consumo (il numero di miglia per gallone, `mpg`, è più elevato), con caratteristiche dimensionali presenti in minore misura (peso, `wt`, cavalli motore, `hp`) e con prestazioni inferiori (il tempo richiesto a percorrere 1/4 di un miglio, `qsec`, è più elevato). La somma dei quadrati all'interno dei gruppi `withinss` dipende dall'omogeneità interna e dalla numerosità del gruppo.

5.6 Discussione

I metodi gerarchici presentano un evidente vantaggio dal punto di vista computazionale; tuttavia risultano maggiormente sensibili agli *outlier* e non consentono di falsificare la configurazione raggiunta: una volta che un'unità è stata attribuita ad un gruppo permane al suo interno per sempre. I metodi non gerarchici non soffrono di questo problema, ma richiedono l'opportuna scelta dei seed.

E’ buona norma applicare una pluralità di metodi per verificare la stabilità dei gruppi: si applica una analisi gerarchica prima per identificare il numero dei gruppi e gli eventuali outlier; si applica poi una tecnica non gerarchica per consentire di modificare la configurazione raggiunta. La determinazione del numero dei cluster può avvenire sulla base dell’informazione a priori o della distanza alla quale avviene l’aggregazione.

Con riferimento all’interpretazione dei raggruppamenti effettuati, il risultato finale dell’analisi dei grappoli è un elenco di unità catalogate a seconda del cluster di appartenenza; al fine di interpretare la configurazione raggiunta si rende necessario tornare alla matrice dei dati di partenza e costruire il profilo medio del gruppo.

Capitolo 6

Analisi delle componenti principali

6.1 Introduzione

Si consideri la situazione prospettata nel grafico 6.1, che rappresenta le coppie dei valori standardizzati del reddito e del consumo pro-capite dei 92 comuni della regione Umbria, stimati con riferimento all'anno 1994 (cfr. [12]). La concentrazione dei punti lungo una direzione principale ben definita indica che le due variabili presentano una correlazione lineare molto elevata e di segno positivo. Supponiamo, a puro titolo illustrativo, di voler pervenire ad un indicatore sintetico che rappresenti il livello delle due grandezze economiche e che minimizzi la perdita di informazione conseguente a tale sintesi. Desideriamo, inoltre, che tale indicatore sia una combinazione lineare delle misurazioni originarie.

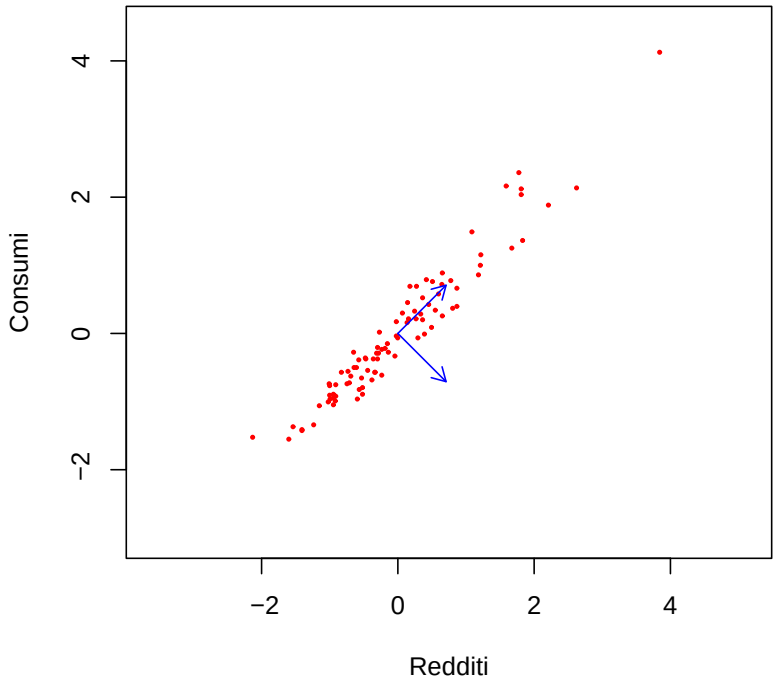
Dal punto di vista geometrico ciò equivale a determinare un sottospazio di dimensione unitaria (una retta nel piano) lungo il quale i punti siano proiettati in modo tale da rappresentare con la massima risoluzione possibile, entro determinati vincoli, le differenze esistenti tra le unità.

L'analisi delle componenti principali, la cui logica verrà esposta nel presente capitolo, consente di determinare l'indicatore richiesto come

$$\text{Indicatore} = 0.71 \cdot \text{Reddito} + 0.71 \cdot \text{Consumo},$$

il quale, pertanto, non differisce dalla somma semplice o dalla media aritmetica semplice delle variabili originarie, se non per un fattore di scala. Il sottospazio generato dalla combinazione lineare è individuato nel grafico dal vettore che si muove nella direzione principale; i punteggi dei comuni con riferimento all'indicatore sono ottenibili proiettando ortogonalmente i punti rappresentativi delle unità lungo questa direzione, la quale, come risulterà dalla trattazione successiva, massimizza la varianza delle proiezioni (si può stabilire formalmente che la nostra combinazione lineare spiega una quota pari al 98% della varianza totale delle misurazioni originarie).

Figura 6.1: Reddito e consumo pro-capite (standardizzati) dei 92 comuni umbri.



La parte dell'informazione di partenza che viene perduta mediante tale sintesi viene misurata dalla proiezione lungo la direzione ortogonale rispetto a quella principale. Essa risulta quantificabile come segue:

$$\text{Residuo} = 0.71 \cdot \text{Reddito} - 0.71 \cdot \text{Consumo}.$$

Si osservi che dalla conoscenza dell'indicatore e del residuo possiamo ricostruire (a meno di un fattore di scala) l'informazione di partenza: aggiungendo e sottraendo il residuo dall'indicatore si ottengono grandezze proporzionali rispettivamente al reddito e al consumo.

In generale, l'analisi delle componenti principali mira a conseguire una riduzione della dimensionalità dell'informazione in presenza di un insieme di variabili fortemente correlate, mediante la definizione di un set di combinazioni lineari delle misurazioni originarie, tra loro incorrelate, ed ordinate in modo tale che la prima componente sintetizza la quota massima possibile della variabilità totale.

Sia $\boldsymbol{x}_i$ un vettore contenente p misurazioni sull'unità i, $i = 1, \ldots, n$, e supponiamo che le misurazioni siano centrate, ovvero

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = \mathbf{0}.$$

La matrice di covarianza delle p misurazioni è

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

Dato un vettore $\boldsymbol{a}$ uscente dall'origine e di lunghezza unitaria, $||\boldsymbol{a}|| = 1$, denotiamo con $\boldsymbol{x}_i^*$ la proiezione ortogonale lungo $\boldsymbol{a}$. Pertanto, applicando la regola del parallelogramma, è possibile individuare un vettore $\boldsymbol{r}_i$, ortogonale a $\boldsymbol{x}_i^*$ ($\boldsymbol{r}_i'\boldsymbol{x}_i^* = 0$) tale che:

$$\boldsymbol{x}_i = \boldsymbol{x}_i^* + \boldsymbol{r}_i \tag{6.1}$$

In virtù dell'ortogonalità, vale la seguente eguaglianza:

$$||\boldsymbol{x}_i||^2 = ||\boldsymbol{x}_i^*||^2 + ||\boldsymbol{r}_i||^2,$$

come è agevole verificare moltiplicando entrambi i membri della (6.1) per $\boldsymbol{x}_i'$. Scrivendo, inoltre,

$$\boldsymbol{x}_i^* = y_i \boldsymbol{a},$$

si ha che y_i rappresenta la coordinata dell'unità i nel sottospazio di proiezione e può essere espressa nei termini del prodotto scalare:

$$y_i = \frac{\boldsymbol{x}_i'\boldsymbol{a}}{||\boldsymbol{a}||^2} = \boldsymbol{x}_i'\boldsymbol{a}, \quad i = 1, \ldots, n.$$

Si noti che le nuove coordinate sono centrate attorno allo zero:

$$\bar{y} = \frac{1}{n}\sum y_i = \left(\frac{1}{n}\sum \boldsymbol{x}_i'\right)\boldsymbol{a} = 0.$$

Ci poniamo ora il problema di determinare il sottospazio di dimensione unitaria in modo tale che la somma dei quadrati degli scarti perpendicolari tra valori osservati ($\boldsymbol{x}_i$) e la loro proiezione ($\boldsymbol{x}_i^*$) sia minima:

$$\min\left\{\sum_{i=1}^{n} ||\boldsymbol{r}_i||^2\right\} = \min\left\{\sum_{i=1}^{n} \boldsymbol{r}_i'\boldsymbol{r}_i\right\}.$$

Ciò equivale a massimizzare la dispersione (varianza) delle proiezioni y_i:

$$\min\left\{\sum_{i=1}^{n} ||\boldsymbol{r}_i||^2\right\} = \min\left\{\sum_{i=1}^{n} ||\boldsymbol{x}_i||^2 - \sum_{i=1}^{n} ||\boldsymbol{x}_i^*||^2\right\} = \max\left\{\sum_{i=1}^{n} ||y_i\boldsymbol{a}||^2\right\}$$

In considerazione della normalizzazione $||\boldsymbol{a}|| = 1$, occorre individuare $\boldsymbol{a}$ in modo da massimizzare la varianza

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}'\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}.$$

6.2 Determinazione delle componenti principali

Nella sezione precedente siamo giunti alla conclusione che al fine di determinare il sottospazio di proiezione occorre risolvere il problema di massimo vincolato:

$$\max\{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}\} \qquad \text{s.v.} \;\; \boldsymbol{a}'\boldsymbol{a} = 1, \tag{6.2}$$

che equivale ad individuare il massimo del lagrangiano

$$\phi(\boldsymbol{a}, \lambda) = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} - \lambda(\boldsymbol{a}'\boldsymbol{a} - 1).$$

Le condizioni del primo ordine (ottenute eguagliando a zero il vettore delle derivate parziali rispetto alle incognite) forniscono:

$$\frac{\partial \phi}{\partial \boldsymbol{a}} = 2\boldsymbol{S}\boldsymbol{a} - 2\lambda\boldsymbol{a} = \boldsymbol{0}$$

$$\frac{\partial \phi}{\partial \lambda} = \boldsymbol{a}'\boldsymbol{a} - 1 = 0,$$

e danno pertanto luogo al sistema omogeneo di p equazioni in $p + 1$ incognite

$$(\boldsymbol{S} - \lambda\boldsymbol{I})\boldsymbol{a} = \boldsymbol{0}$$

dove $\boldsymbol{a}$ soddisfa il vincolo di normalizzazione $\boldsymbol{a}'\boldsymbol{a} = 1$. Il problema coincide pertanto con quello di determinare gli autovalori e gli autovettori della matrice di covarianza, $\boldsymbol{S}$. In particolare, λ è l'autovalore più grande della matrice $\boldsymbol{S}$, la quale, peraltro, risultando semidefinita positiva, avrà p autovalori non negativi, mentre $\boldsymbol{a}$ rappresenta l'autovettore corrispondente all'autovalore massimo (normalizzato in modo da avere $\boldsymbol{a}'\boldsymbol{a} = 1$.

La variabile di coordinate $y_i = \boldsymbol{x}_i'\boldsymbol{a}$ viene detta *componente principale* ed è determinata in modo da estrarre la quota massima di varianza dalle misurazioni originarie. In termini matriciali, denotando con $\boldsymbol{y}$ il vettore $n \times 1$ contenente i valori della componente principale per le n unità (y_i),

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a}$$

Si noti che $\boldsymbol{S}\boldsymbol{a} = \lambda\boldsymbol{a}$ implica che la varianza di detta componente è pari λ; infatti

$$\mathsf{Var}(\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} = \lambda\boldsymbol{a}'\boldsymbol{a} = \lambda.$$

6.3 Autovalori e autovettori

Data la matrice $\boldsymbol{S}$, quadrata di dimensione p e simmetrica, consideriamo il problema di determinare uno scalare λ e un vettore $\boldsymbol{a}$ che soddisfano il sistema di equazioni:

$$\boldsymbol{S}\boldsymbol{a} = \lambda\boldsymbol{a}$$

Si noti che il problema è indeterminato, dal momento che le incognite sono $p + 1$: gli elementi del vettore $\boldsymbol{a}$ e lo scalare λ. Quest'ultimo è detto autovalore o valore caratteristico (latente) della matrice $\boldsymbol{S}$, mentre $\boldsymbol{a}$ è denominato autovettore o vettore caratteristico (latente). A tale indeterminatezza si pone rimedio imponendo il vincolo di normalizzazione

$$\boldsymbol{a}'\boldsymbol{a} = 1,$$

mediante il quale si richiede che l'autovettore abbia lunghezza unitaria.

Riscrivendo il sistema nella forma:

$$(\boldsymbol{S} - \lambda\boldsymbol{I})\boldsymbol{a} = \boldsymbol{0},$$

si evidenzia che, per dato λ, il sistema è omogeneo ed ammette una soluzione non banale ($\boldsymbol{a} \neq 0$) se e solo se

$$|\boldsymbol{S} - \lambda\boldsymbol{I}| = 0.$$

Questa condizione fornisce un'equazione in λ di grado p, detta equazione caratteristica, che sarà appunto utilizzata per determinare λ. Sostituendo a turno ciascuna delle p soluzioni in $(\boldsymbol{S} - \lambda\boldsymbol{I})\boldsymbol{a} = \boldsymbol{0}$, denotate $\lambda_1, \ldots, \lambda_p$, si determinano in corrispondenza gli autovettori $\boldsymbol{a}_h, h = 1, \ldots, p$, risolvendo il sistema omogeneo di p equazioni in p incognite.

Ora, è possibile dimostrare che:

1. Gli autovalori di una matrice simmetrica sono reali (nel caso generale possono essere complessi); questi possono essere distinti o presentarsi ripetuti più volte (molteplicità)

2. Gli autovettori corrispondenti ad autovalori distinti sono ortogonali: siano $\boldsymbol{a}_h$ e $\boldsymbol{a}_k$ due autovettori corrispondenti alle radici λ_h e $\lambda_k \neq \lambda_h$; allora, $\boldsymbol{a}_h'\boldsymbol{a}_k = 0$. Inoltre, se un autovalore ha molteplicità m, esistono in corrispondenza m autovettori ortogonali.

3. La proprietà precedente abbinata al vincolo di normalizzazione ($\boldsymbol{a}'\boldsymbol{a} = 1$) implica che gli autovettori di una matrice simmetrica costituiscono un insieme ortonormale:

$$\boldsymbol{a}_h'\boldsymbol{a}_k = \begin{cases} 1 & h = k \\ 0 & h \neq k \end{cases}$$

   Raccogliendo i p autovettori nella matrice $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p]$, si ha $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$ e $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$, ovvero $\boldsymbol{A}$ è una matrice ortogonale (l'inversa e la trasposta coincidono).

6.3.1 Scomposizione spettrale di una matrice

I p sistemi $\boldsymbol{S}\boldsymbol{a}_h = \lambda_h\boldsymbol{a}_h, h = 1, \ldots, p$, possono essere raccolti in

$$\boldsymbol{S}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{L},$$

dove $\boldsymbol{L} = \text{diag}(\lambda_1, \ldots, \lambda_p)$. Premoltiplicando entrambi i membri per $\boldsymbol{A}'$, si ottiene

$$\boldsymbol{A}'\boldsymbol{S}\boldsymbol{A} = \boldsymbol{L},$$

da cui l'affermazione che la matrice degli autovettori diagonalizza $\boldsymbol{S}$.

Premoltiplicando l'espressione precedente per $\boldsymbol{A}$ e postmoltiplicando per $\boldsymbol{A}'$, si consegue la scomposizione spettrale della matrice (simmetrica) $\boldsymbol{S}$:

$$\boldsymbol{S} = \boldsymbol{A}\boldsymbol{L}\boldsymbol{A}' = \sum_{h=1}^{p} \lambda_h \boldsymbol{a}_h \boldsymbol{a}_h'$$

Si noti che il rango di $\boldsymbol{A}$ è pieno (dal momento che $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$), e questo implica che

$$\text{rank}(\boldsymbol{S}) = \text{rank}(\boldsymbol{L})$$

Pertanto, il rango di una matrice simmetrica è pari al numero di autovalori non nulli. Se una matrice ha uno o più autovalori nulli, allora si ha $\boldsymbol{S}\boldsymbol{a} = \boldsymbol{0}$ il che implica che la matrice è singolare (rango ridotto).

Inoltre, è immediato mostrare che il determinante di una matrice è uguale al il prodotto degli autovalori:

$$|\boldsymbol{S}| = |\boldsymbol{A}\boldsymbol{L}\boldsymbol{A}'| = |\boldsymbol{A}'|\,|\boldsymbol{L}|\,|\boldsymbol{A}| = |\boldsymbol{A}'\boldsymbol{A}|\,|\boldsymbol{L}| = |\boldsymbol{L}| = \prod_{h=1}^{p} \lambda_h,$$

mentre la traccia è pari alla somma degli autovalori:

$$\text{tr}(\boldsymbol{S}) = \sum_{h=1}^{p} \lambda_h$$

6.3.2 Esempi illustrativi

Supponiamo di disporre di due variabili standardizzate caratterizzate dalla matrice di covarianza (correlazione):

$$\boldsymbol{S} = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}.$$

Al fine di determinare gli autovalori si risolve l'equazione caratteristica:

$$|\boldsymbol{S} - \lambda \boldsymbol{I}| = \lambda^2 - 2\lambda + 0.36 = 0,$$

che fornisce un'equazione di secondo grado, le cui soluzioni sono $\lambda_1 = 1.8$ e $\lambda_2 = 0.2$. L'autovettore $\boldsymbol{a}_1 = [a_{11}, a_{21}]'$ viene ottenuto a soluzione del sistema omogeneo di due equazioni:

$$(\boldsymbol{S} - 1.8\boldsymbol{I})\boldsymbol{a}_1 = \begin{bmatrix} -.8 & .8 \\ .8 & -.8 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

per il quale, $a_{11} = a_{21}$, ovvero, $\boldsymbol{a}_1 = a_{21}\boldsymbol{i}_2$. Si osservi che l'autovettore è determinato a meno di un fattore scalare e che è proporzionale ad un vettore unitario. Al fine di determinare una unica soluzione entra in gioco il vincolo di normalizzazione, $\boldsymbol{a}_1'\boldsymbol{a}_1 = a_{11}^2 + a_{21}^2 = 1$, da cui consegue $2a_{21}^2 = 1$, ovvero, $a_{21} = 1/\sqrt{2}$. In conclusione, $\boldsymbol{a}_1 = (1/\sqrt{2})\boldsymbol{i} = [.71, .71]'$ (si noti la similarità con l'esempio riferito ai redditi ed ai consumi riportato nella prima sezione di questo capitolo).

Analogamente, si determina il secondo autovettore in corrispondenza dell'autovalore $\lambda_2 = .2$:

$$\boldsymbol{a}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

La scomposizione spettrale della matrice $\boldsymbol{S}$ è dunque

$$\boldsymbol{S} = 1.8\frac{1}{2}\boldsymbol{i}\boldsymbol{i}' + 0.2\frac{1}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}[1 \quad -1];$$

il primo addendo costituisce un'approssimazione di rango 1 della matrice; sostituire l'informazione di partenza con la prima componente principale equivale ad interpolare $\boldsymbol{S}$ con una matrice di covarianza con elementi tutti pari a 0.9. Questo determina una sottostima nella rappresentazione delle varianze pari a 0.1 e una sovrastima della covarianza dello stesso ammontare.

Il calcolo degli autovalori e degli autovettori in R viene effettuato dalla funzione `eigen( )`:

```
> S <- matrix(c(1,.8,.8,1),2)
> S
     [,1] [,2]
[1,]  1.0  0.8
[2,]  0.8  1.0
> eigen(S)
$values
[1] 1.8 0.2

$vectors
          [,1]       [,2]
[1,] 0.7071068  0.7071068
[2,] 0.7071068 -0.7071068
```

La traccia della matrice $\boldsymbol{S}$ è pari a $2 = \lambda_1 + \lambda_2$, mentre il determinante è uguale a $1.8 \cdot 0.2 = 0.36$. R non contiene una funzione specifica per il calcolo del determinante di una matrice; a tal fine possiamo introdurne una fondata sul prodotto degli autovalori ottenuti come output della funzione `eigen`:

```
det <- function(S)
       {  if (ncol(S) != nrow(S)) print("Attenzione: matrice non quadrata")
          else prod(eigen(S)$values)
       }
```

Nel caso di variabili incorrelate ed eteroschedastiche, $\boldsymbol{S} = \text{diag}(s_1^2, s_2^2, \ldots, s_p^2)$, dove senza perdita di generalità assumiamo l'ordinamento $s_1^2 \geq s_2^2 \geq \cdots \geq s_p^2$, è immediato mostrare che gli autovalori sono pari alle varianze ($\lambda_h = s_h^2, h = 1, \ldots, p$, dal momento che l'equazione caratteristica è $(s_1^2 - \lambda)(s_2^2 - \lambda) \cdots (s_p^2 - \lambda)$) e gli autovettori sono i vettori canonici $\boldsymbol{e}_h = [0, \ldots, 0, 1, 0, \ldots, 0]'$.

6.4 La soluzione generale

Supposto che gli autovalori della matrice $\boldsymbol{S}$ siano ordinati in senso non crescente,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0,$$

la prima componente principale, $\boldsymbol{y}_1$, il cui elemento generico è y_{i1}, viene definita come combinazione lineare delle p variabili originarie in $\boldsymbol{X}$, con pesi forniti dagli elementi del primo autovettore (corrispondente all'autovalore più grande, λ_1): $\boldsymbol{y}_1 = \boldsymbol{X}\boldsymbol{a}_1$. La seconda componente, i cui pesi sono contenuti nel vettore $\boldsymbol{a}_2$, massimizza la varianza residua sotto il vincolo di ortonormalità, vale a dire:

$$\max\{\boldsymbol{a}_2'(\boldsymbol{S} - \lambda_1 \boldsymbol{a}_1 \boldsymbol{a}_1')\boldsymbol{a}_2\} \quad \text{s.v } \boldsymbol{a}_2'\boldsymbol{a}_1 = 0, \quad \boldsymbol{a}_2'\boldsymbol{a}_2 = 1,$$

La soluzione di questo problema consente di determinare $\boldsymbol{a}_2$ come l'autovettore corrispondente all'autovalore λ_2 (il secondo in ordine di grandezza), il quale coincide con la varianza della componente medesima. Possiamo continuare a determinare le rimanenti componenti seguendo la medesima logica, giungendo fino all'ultima, $\boldsymbol{y}_p = \boldsymbol{X}\boldsymbol{a}_p$, la quale ha la varianza più piccola, ed i coefficienti di $\boldsymbol{a}_p$ sono forniti dall'autovettore corrispondente a λ_p.

In generale, miriamo ad ottenere una rappresentazione della matrice $\boldsymbol{X}$ in un sottospazio ortogonale (iperpiano) a $r < p$ dimensioni. Le coordinate dei punti nel sottospazio di proiezione sono contenute in una matrice $\boldsymbol{Y}$, di dimensione $n \times r$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}_r$$

e sono ottenute come combinazione lineare delle coordinate iniziali. I coefficienti della combinazione lineare sono contenuti nella matrice $\boldsymbol{A}_r$, di dimensione $p \times r$

$$\boldsymbol{A}_r = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ \cdots \ \boldsymbol{a}_r]$$

e sono detti *loadings*, poiché forniscono il peso assegnato alle variabili originarie nella definizione delle componenti principali; le colonne di $\boldsymbol{A}_r$ costituiscono un set ortonormale ($\boldsymbol{A}_r'\boldsymbol{A}_r = \boldsymbol{I}_r$).

Il punteggio sull'h-esima componente principale per l'unità i è fornito da

$$y_{ih} = a_{1h}x_{i1} + a_{2h}x_{i2} + \ldots + a_{ph}x_{ip} = \boldsymbol{a}_h'\boldsymbol{x}_i$$

e, sommando per $i = 1, \ldots, n$ e dividendo per n,

$$\bar{y}_h = a_{1h}\bar{x}_1 + a_{2h}\bar{x}_2 + \ldots + a_{ph}\bar{x}_p = \boldsymbol{a}_h'\bar{\boldsymbol{x}}.$$

Inoltre, denotando

$$\boldsymbol{y}_h = \boldsymbol{X}\boldsymbol{a}_h$$

la varianza dell'h-esima componente risulta,

$$\frac{1}{n}\sum_{i=1}^{n}(y_{ih} - \bar{y}_h)^2 = \boldsymbol{a}_h'\boldsymbol{S}\boldsymbol{a}_h = \lambda_h$$

Infine, la covarianza tra l'h-esima componente e la k-esima componente risulta pari a zero.

In termini matriciali, possiamo riassumere i risultati con le seguenti espressioni:
a) definizione delle componenti:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}_r$$

b) vettore delle r medie delle c.p.:

$$\bar{\boldsymbol{y}} = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{i}_n = \frac{1}{n}\boldsymbol{A}_r'\boldsymbol{X}'\boldsymbol{i}_n = \boldsymbol{A}_r'\bar{\boldsymbol{x}}$$

c) matrice di covarianza delle r c.p.:

$$\frac{1}{n}\boldsymbol{Y}'\boldsymbol{Y} - \bar{\boldsymbol{y}}\bar{\boldsymbol{y}}' = \boldsymbol{L}_r = \text{diag}(\lambda_1, \ldots .\lambda_r)$$

6.5 La standardizzazione delle variabili

Se le p misurazioni di partenza sono espresse su unità di misura molto diverse, comportando una notevole differenziazione nelle varianze, può essere consigliabile effettuare una standardizzazione delle misurazioni originarie:

$$x_{ik} \longrightarrow z_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$$

ovvero

$$\boldsymbol{X} \longrightarrow \boldsymbol{Z} = (\boldsymbol{X} - \boldsymbol{i}\bar{\boldsymbol{x}}')\boldsymbol{D}^{-1/2}$$

dove $\boldsymbol{D} = \text{diag}(s_1^2, \ldots, s_p^2)$. La matrice di correlazione può essere pertanto espressa:

$$\boldsymbol{R} = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}$$

Le componenti principali saranno definite come combinazione lineare di $\boldsymbol{Z}$ con *loadings* $\boldsymbol{A}_r$ che sono forniti dalla scomposizione spettrale della matrice di correlazione ($\boldsymbol{R} = \boldsymbol{ALA}'$):

$$\boldsymbol{Y} = \boldsymbol{ZA}_r$$

ed avranno media nulla e matrice di covarianza $\boldsymbol{L}_r$, dove $\boldsymbol{L}_r$ contiene gli autovalori in ordine decrescente della matrice di correlazione.

Dimostriamo ora che la distanza di Mahalanobis equivale alla distanza euclidea calcolata sulle componenti principali standardizzate.

$$\begin{aligned} {}_M d_{ij}^2 &= (\boldsymbol{z}_i - \boldsymbol{z}_j)' \boldsymbol{R}^{-1} (\boldsymbol{z}_i - \boldsymbol{z}_j) \\ &= (\boldsymbol{z}_i - \boldsymbol{z}_j)' \boldsymbol{A} \boldsymbol{L}^{-1} \boldsymbol{A}' (\boldsymbol{z}_i - \boldsymbol{z}_j) \\ &= (\boldsymbol{y}_i - \boldsymbol{y}_j)' \boldsymbol{L}^{-1} (\boldsymbol{y}_i - \boldsymbol{y}_j) \\ &= (\boldsymbol{y}_i - \boldsymbol{y}_j)' \boldsymbol{L}^{-1/2} \boldsymbol{L}^{-1/2} (\boldsymbol{y}_i - \boldsymbol{y}_j) \\ &= (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_j)' (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_j) \end{aligned}$$

dal momento che $\boldsymbol{y}_i = \boldsymbol{A}' \boldsymbol{z}_i$ e $\tilde{\boldsymbol{y}}_i = \boldsymbol{L}^{-1/2} \boldsymbol{y}_i$ denota le c.p. standardizzate per l'i-esima unità.

Se si utilizzano soltanto r componenti principali si otterrà un'approssimazione alla distanza di Mahalanobis.

6.6 L'analisi delle CP come metodo di proiezione

Dati n punti in uno spazio p dimensionale,

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$$

l'analisi delle componenti principali mira ad ottenere uno spazio di rappresentazione di dimensione ridotte ($r < p$).

Le proprietà caratteristiche della soluzione delle componenti principali sono essenzialmente tre:

1. Proiezione ortogonale: i punti $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$, sono proiettati ortogonalmente nel sottospazio (iperpiano) definito dalle componenti principali, per ottenere $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_i^*, \ldots, \boldsymbol{x}_n^*$, dove $\boldsymbol{x}_i^* = \boldsymbol{A}_r \boldsymbol{y}_i$. Le coordinate dei punti nel nuovo spazio, definito dalle colonne di $\boldsymbol{A}_r$, sono:

   $$\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_i, \ldots, \boldsymbol{y}_n,$$

   con $\boldsymbol{y}_i = \boldsymbol{A}_r' \boldsymbol{x}_i$. Si osservi che $\boldsymbol{x}_i^* = \boldsymbol{A}_r \boldsymbol{A}_r' \boldsymbol{x}_i$; ora, per effetto della ortogonalità della matrice $\boldsymbol{A}$, $\boldsymbol{AA}' = \boldsymbol{I}$, e partizionando $\boldsymbol{A} = [\boldsymbol{A}_r \quad \tilde{\boldsymbol{A}}]$, dove $\tilde{\boldsymbol{A}}$ è una matrice $p \times (p - r)$ che contiene i rimanenti autovettori, si può scrivere:

   $$\boldsymbol{x}_i = \boldsymbol{AA}' \boldsymbol{x}_i = [\boldsymbol{A}_r \;\; \tilde{\boldsymbol{A}}] \begin{bmatrix} \boldsymbol{A}_r' \\ \tilde{\boldsymbol{A}}' \end{bmatrix} = \boldsymbol{A}_r \boldsymbol{A}_r' \boldsymbol{x}_i + \tilde{\boldsymbol{A}} \tilde{\boldsymbol{A}}' \boldsymbol{x}_i = \boldsymbol{x}_i^* + \boldsymbol{r}_i,$$

dove $\boldsymbol{r}_i$ rappresenta il residuo di proiezione, vale a dire la parte di informazione che viene persa sull'unità i per effetto della sintesi effettuata dalle prime r componenti principali.

2. L'iperpiano di proiezione, generato dalle colonne della matrice $\boldsymbol{A}_r$, è orientato in modo da rendere massima la dispersione degli n punti

$$\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_i^*, \ldots, \boldsymbol{x}_n^*.$$

3. L'iperpiano di proiezione è tale da rendere minima la funzione V

$$V = \sum_i \sum_j (d_{ij}^2 - \hat{d}_{ij}^2)$$

dove d_{ij} rappresenta la distanza euclidea tra $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$ nello spazio di partenza e $\hat{d}_{ij}$ rappresenta la distanza euclidea tra le due unità nello spazio di proiezione generato da $\boldsymbol{A}_r$.

Al fine di dimostrare l'ultima proprietà, si assuma, per semplicità, che $\bar{\boldsymbol{x}} = \boldsymbol{0}$, e scriva $d_{ij}^2 = \boldsymbol{x}_i'\boldsymbol{x}_i + \boldsymbol{x}_j'\boldsymbol{x}_j - 2\boldsymbol{x}_i'\boldsymbol{x}_j$; sommando rispetto ad i e j,

$$\sum_i \sum_j d_{ij}^2 = 2n \sum_i \boldsymbol{x}_i'\boldsymbol{x}_i;$$

un risultato analogo vale per $\hat{d}_{ij}^2$:

$$\sum_i \sum_j \hat{d}_{ij}^2 = 2n \sum_i \boldsymbol{x}_i^{*'}\boldsymbol{x}_i^*.$$

Pertanto,

$$V = 2n(\sum_i \boldsymbol{x}_i'\boldsymbol{x}_i - \sum_i \boldsymbol{x}_i^{*'}\boldsymbol{x}_i^*)$$

e $\min V$ equivale a

$$\max \sum_i ||\boldsymbol{x}_i^*||^2$$

6.7 Determinazione del numero delle componenti principali

Dal momento che le componenti principali sono incorrelate e hanno varianza λ_h, la varianza totale delle p componenti principali risulta

$$\sum_{h=1}^{p} s_h^2 = \text{tr}(\boldsymbol{S}) = \sum_{h=1}^{p} \lambda_h.$$

Ai fini della scelta di r, vale a dire del numero minimo di componenti principali sufficienti per ottenere una descrizione sintetica della matrice dei dati, $\boldsymbol{X}$, possiamo utilizzare tre criteri:

- Fissare un limite inferiore, q (ad esempio, $q = .9$), alla quota di varianza spiegata dalle prime r componenti, Q_r,

  $$Q_r = \frac{\sum_{h=1}^{r} \lambda_h}{\sum_{h=1}^{p} \lambda_h}$$

  scegliendo r in modo tale che $Q_r \geq q$.

- Il grafico degli autovalori rispetto al numero d'ordine della componente viene denominato *scree plot*; si sceglie r in corrispondenza del quale il grafico presenta un gomito (*elbow*).

- Criterio di Kaiser: si calcola la media delle varianze, ovvero l'autovalore medio

  $$\bar{\lambda} = \frac{1}{p} \sum_{h=1}^{p} \lambda_h.$$

  Per dati sferici, vale a dire incorrelati e a varianza costante, $\lambda_h = \bar{\lambda}$ (infatti, $\boldsymbol{S} = s^2 \boldsymbol{I}$ e $|\boldsymbol{S} - \lambda \boldsymbol{I}| = (s^2 - \lambda)^p$). Si estraggono le prime r componenti la cui varianza supera tale media, ovvero r è il più grande valore di h tale che $\lambda_h > \bar{\lambda}$.

Se si è proceduto alla standardizzazione delle variabili originarie, la varianza totale è

$$\text{tr}(\boldsymbol{R}) = p,$$

per cui il criterio di Kaiser equivale a scegliere un numero di componenti pari al numero di autovalori superiori all'unità.

6.8 Illustrazione

Il data set `mdspref.dat` contiene i punteggi medi attribuiti su una scala da 1 a 7 su otto attributi relativi a 10 bibite.

```
> drinks <- read.table("mdspref.dat",header=T)
> X <- t(drinks)
> library(mva)
> cp.drinks <- princomp(X)
> summary(cp.drinks)
```

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4      Comp.5      Comp.6
Standard deviation     2.5006260 1.7383913 0.57814153 0.45368404 0.311120966 0.242115128
Proportion of Variance 0.6259323 0.3024997 0.03345786 0.02060330 0.009689211 0.005867779
Cumulative Proportion  0.6259323 0.9284320 0.96188984 0.98249314 0.992182353 0.998050132
                            Comp.7       Comp.8
Standard deviation     0.131628571 0.0464037630
Proportion of Variance 0.001734324 0.0002155442
Cumulative Proportion  0.999784456 1.0000000000
```

L'output presentato da R fornisce le deviazioni standard delle componenti, pari a $\sqrt{\lambda_h}$, la quota di varianza spiegata da ciascuna componente, $\lambda_h / \sum_k \lambda_k$ e la quota cumulata spiegata dalle prime h componenti, Q_h. L'analisi mostra che le prime due componenti spiegano una quota pari al 93% della varianza totale.

```
> loadings(cp.drinks)[,1:2]
                Comp.1      Comp.2
Fruity      -0.3810421  0.71127292
Carbonation  0.2939139  0.03123506
Calories    -0.1870931 -0.17416675
Tart         0.3667315  0.31992189
Thirst       0.1246746  0.51200424
Popularity   0.5418241 -0.07561199
Aftertaste   0.2442897  0.30260352
Pick-up      0.4797200  0.03157099
par(mfrow=c(2,2))
screeplot(cp.drinks, type="lines", main = "Screeplot")
barplot(loadings(cp.drinks)[,1], cex = .6, main = "Pesi 1a CP")
barplot(loadings(cp.drinks)[,2], cex = .6, main = "Pesi 2a CP")
plot(cp.drinks$scores[,1:2], type="n", xlim=c(-4,4), ylim=c(-4,4),
     main = "Grafico delle prime due CP")
text(cp.drinks$scores[,1:2], dimnames(X)[[1]], cex=.6)
```

La prima componente è negativamente correlata con gli attributi `Fruity` e `Calories` e positivamente con tutti gli altri; per tale motivo, i prodotti ipocalorici e dietetici, per i quali i due attributi sono presenti in minore misura, hanno un punteggio positivo su questa componente. Se si rovesciasse il punteggio (prendendo il complemento a 7) per `Fruity` e `Calories`, la prima componente definisce una combinazione lineare degli 8 attributi con pesi tutti positivi, e non sarebbe molto distante da una media aritmetica. La seconda componente ha un peso molto elevato su `Fruity` e discrimina le bibite prevalentemente su questo attributo.

Il grafico a dispersione dei punteggi (centrati) delle prime due componenti (cfr. figura 6.2) mette in luce la similarità tra le bibite `CokeCl`, `Coke` e `Pepsi`.

L'analisi delle componenti principali può essere replicata utilizzando la scomposizione spettrale della matrice di covarianza:

```
S <- cov(X)
ds <- eigen(S)
lambda <- ds$values
A <- ds$vectors
scores <- X %*% A
# NB cp.drinks$score sono ottenuti centrando le componenti principali
scale(scores, scale=F)
```

Figura 6.2: Analisi delle componenti principali per il data set `mdspref.dat`.