5. Analisi dei Gruppi (Cluster Analysis)

Cosa è l'analisi dei gruppi?

- Viene utilizzata per classificare rispondenti in gruppi omogenei detti clusters.
- Esamina relazioni di interdipendenza: nessuna distinzione tra variabile dipendente ed indipendenti
- Obiettivo: classificare unità statistiche (individui) in gruppi omogenei in base alle variabili considerate. All'interno del gruppo le unità dovrebbero essere omogenee.
- Dati: a seconda della scala di misura che caratterizza le variabili si devono utilizzare funzioni di distanza/similarità diverse.

Differenze tra l'analisi cluster e le altre tecniche multivariate

- Nell'analisi cluster non è necessario avere nessuna indicazione a priori sul gruppo di appartenenza di ogni singola unità.

- Nell’analisi discriminante per definire una regola classificazione è necessario stabilire a priori l’appartenenza di un’unità ad un gruppo.

- L’analisi cluster è una tecnica per la riduzione del numero di unità
- L’analisi fattoriale è una tecnica per la riduzione del numero delle variabili e si basa sull’analisi delle relazioni tra variabili

Zoom sul confronto analisi fattoriale e analisi cluster

L’analisi fattoriale assume che le relazioni tra le variabili inserite nel modello di analisi siano lineari (si basa sulla matrice di varianze covarianze), mentre la forma delle relazioni tra le variabili è trascurabile nell’analisi dei cluster. Questo non esclude che le due tecniche possono portare a conclusioni analoghe: dopo aver eseguito un’analisi cluster si possono individuare le variabili più discriminanti tra le unità; e dopo un’analisi fattoriale si possono individuare le unità più simili rispetto ai fattori individuati.
I due metodi vengono impiegati in sequenza per ottenere effetti retroattivi, nel senso che (Rizzi,

1985) “dal concetto di gruppo si prendono le mosse per l’individuazione delle caratteristiche del fenomeno, le quali a loro volta permettono una migliore individuazione del gruppo stesso”.
Più precisamente:

1. la procedura di raggruppare prima le unità statistiche e poi di effettuare la ricerca dei fattori in ogni gruppo di dimensione sufficiente è indicata al posto dell’analisi fattoriale sull’intero insieme di unità quando i valori delle correlazioni globali sono modesti mentre quelli delle correlazioni interne sono rilevanti;
2. quando il numero di variabili è troppo grande per applicare una tecnica di raggruppamento, e in generale, quando si desideri eliminare in modo mirato la ridondanza nei dati osservati. In questo modo si riducono le informazioni ricorrendo ai fattori e si esegue un’analisi cluster sui punteggi fattoriali

Concetto di similarità per la formazione dei clusters

- Le unità all'interno dello stesso clusters dovrebbero essere simili tra loro ma differenti dalle unità appartenenti ad altri clusters.
- La situazione ideale sarebbe che una unità appartenesse ad uno ed un solo cluster e che tutti i cluster fossero disgiunti
- In realtà i confini di ogni singolo cluster non sono ben definiti
- Le procedure che utilizziamo assegnano una unità ad uno ed un solo cluster
- Il numero di cluster che la procedura definisce può essere molto ampio, l'algoritmo dovrebbe produrre il miglior raggruppamento.

Applicazioni dell'analisi cluster nelle ricerche di mercato

- ❖ Segmentazione del mercato: si formano gruppi di consumatori in base ai benefici che trovano nell'acquisto di un prodotto.
- ❖ Comportamento dei consumatori: si identificano gruppi omogenei di consumatori e si esaminano i comportamenti d acquisto separatamente.
- ❖ Sviluppo e ricerca di opportunità per potenziali nuovi prodotti:
  - cluster di prodotti per identificare prodotti competitivi nel mercato (nicchie di competitività)
  - marche appartenenti allo stesso cluster possono essere con maggiore probabilità concorrenti rispetto a marche appartenenti a cluster diversi.
- ❖ Selezione di mercati (aree test di mercato): gruppi di aree (città) omogenee, in modo da generalizzare i risultati ottenuti in un'area alle rimanenti aree dello stesso cluster, riducendo il numero complessivo di aree test.
- ❖ Riduzione dei dati:

- strumento generale di riduzione dei dati in modo da rendere più gestibili numerosità molto elevate di osservazioni
- condurre analisi diverse su clusters diversi

Come si conduce un'analisi cluster

1. Si formula il problema identificando le variabili di classificazione

2. Si seleziona una funzione di distanza tra le unità

3. Si seleziona una procedura di "raggruppamento"

4. Si decide il numero di clusters

5. Interpretazione dei cluster

6. Validità dei raggruppamenti individuati

1. Si formula il problema

- Si selezionano le variabili alla base della procedura di raggruppamento (molto importante).
- L'inclusione anche di una sola variabile irrilevante può modificare i risultati
- Le variabili selezionate dovrebbero essere dei buoni indicatori delle "similarità" tra le unità (similarità rilevanti per il problema oggetto di studio)

Statements dell'esempio

V_1: Shopping is fun.
V_2: Shopping is bad for your budget.
V_3: I combine shopping with eating out.
V_4: I try to get the best buys while shopping.
V_5: I don't care about shopping.
V_6: You can save a lot of money by comparing prices.

2. Si sceglie una funzione di distanza/similarità

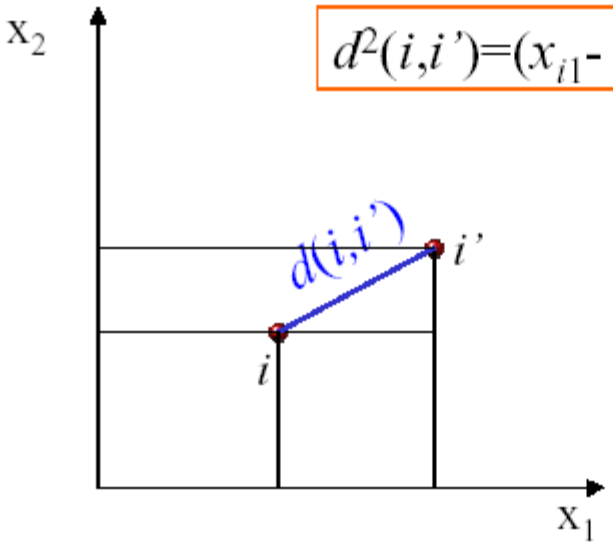
Una funzione di distanza tra due unità i, i' soddisfa le seguenti proprietà

1. $d(i,i') \geq 0$ non negatività
2. $d(i,i') = d(i',i)$ simmetria
3. $d(i,i') = 0$ sse i e i' hanno le stesse x
4. $d(i,i') > d(i,i'')$ ➔ i più vicina a i''

Esistono infinite possibilità per definire funzioni di distanza, per un'analisi approfondita (argomento facoltativo) si veda *Troilo (2003), Fabbris (1997).*Tuttavia la funzione di distanza (per dati quantitativi) più utilizzata è sicuramente la distanza Euclidea.

Distanza euclidea: caso bivariato p=2

$$d^2(i,i') = (x_{i1} - x_{i',1})^2 + (x_{i2} - x_{i',2})^2$$



Distanza euclidea

$$d(i,i') = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2}$$

$$d(i,i')^2 = \left(\mathbf{x}_i - \mathbf{x}_{i'}\right)' \left(\mathbf{x}_i - \mathbf{x}_{i'}\right)$$

$\mathbf{x}_i=(x_{i1}, \ldots, x_{ip})'$ è il vettore delle covariate X corrispondenti alla i^{ma} unità statistica

Combina scarti tra variabili che possono essere espresse in unità di misura diverse

3. Si seleziona una procedura di “raggruppamento”

Le tecniche per il raggruppamento si possono dividere in due grandi categorie.
Analisi gerarchica dei gruppi: in questo caso ogni gruppo fa parte di una classe più ampia, la quale è contenuta a sua volta in una classe di ampiezza superiore, e così via fino al gruppo che contiene tutte le unità.
Analisi non gerarchiche: sono tecniche che generano gruppi non gerarchizzabili; in questo caso si deve decidere a priori il numero di gruppi e le soluzioni con G o G-1 gruppi sono confrontabili solo attraverso indici sintetici e non gruppo per gruppo.

Per la trattazione delle tecniche specifiche (argomento facoltativo) si veda *Troilo (2003), Fabbris (1997).*

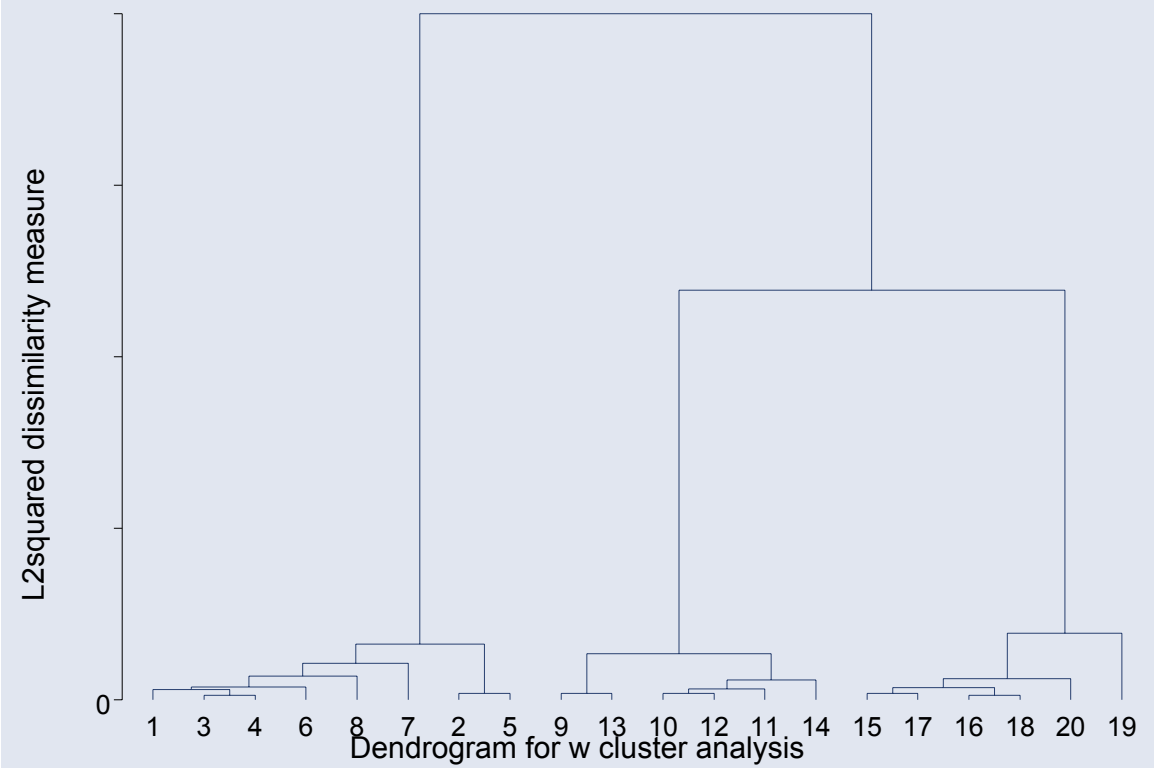
Tutte le specifiche tecniche gerarchiche e non presentano pregi e difetti, molto spesso la scelta si basa anche sull’efficienza computazionale della tecnica.

```
/*analisi cluster*/
/*item v1-v6 attitudini verso lo shopping*/

use "F:\written\didattica\statistica per le analisi di mercato\dati/shopping"

*clusterizzazione gerarchica (metodo ward)

cluster wardslinkage v1-v6, name(w)
cluster dendrogram
```



/*le linee verticali del dendrogramma segnalano l'unione di due cluster, mentre le posizioni di tali linee indicano le distanze alle quali tali cluster vengono aggregati: in questo caso i tre cluster sembrano abbastanza delineati*/

/*In STATA esistono anche altri due indicatori per la scelta del numero dei cluster: il Calinski Harabasz pseudo F statistics e l'indicatore Duda and Hart.
Per entrambi gli indicatori, valori grandi indicano cluster più distinti*/

```
/*valori grandi indicano quando fermarsi*/
cluster stop , rule(calinski) groups(2/4)
```

Number of clusters	Calinski/ Harabasz pseudo-F
2	16.26
3	26.56
4	21.84

/*il valore più elevato si ottiene per 3 cluster*/

```
cluster gen cl3=groups(3), name (w)

sort cl3 cod
forvalues i=1/3  {
         disp "cluster `i'"
         list cod if cl3==`i'
                  }
```

/*Andiamo ad elencare le unità appartenenti ai singoli cluster*/

cluster 1

	cod
1.	1
2.	3
3.	6
4.	7
5.	8
6.	12
7.	15
8.	17

cluster 2

	cod
9.	2
10.	5
11.	9
12.	11
13.	13
14.	20

cluster 3

```
     +-----+
     | cod |
     |-----|
 15. |   4 |
 16. |  10 |
 17. |  14 |
 18. |  16 |
 19. |  18 |
 20. |  19 |
     +-----+
```

/*Vediamo i valori medi degli indicatori nei singoli cluster*/

tabstat v1-v6, by(cl3)

```
Summary statistics: mean
  by categories of: cl3

     cl3 |        v1        v2        v3        v4        v5        v6
---------+------------------------------------------------------------
       1 |      5.75     3.625         6     3.125     1.875     3.875
       2 |  1.666667         3  1.833333       3.5       5.5  3.333333
       3 |       3.5  5.833333  3.333333         6       3.5         6
---------+------------------------------------------------------------
   Total |      3.85       4.1      3.95       4.1      3.45      4.35
------------------------------------------------------------------------
```

```
/*A questo punto cerchiamo di comprendere quali sono le caratteristiche dei
singoli cluster*/
/*
```

CL1: V1 V3 (shopping 'spendaccione')

V1: Shopping is fun.
V3: I combine shopping with eating out.

CL2 V5 (disinteressati allo shopping)

V5: I don't care about shopping.

CL3 V2 V4 V6 (Shopping in economia)

V2: Shopping is bad for your budget.
V4: I try to get the best buys while shopping.
V6: You can save a lot of money by comparing prices.*/

/*Ripetiamo lo stesso esempio utilizzando un metodo non gerarchico, il metodo delle k medie: in questo caso devosempre specificare il numero di cluster che voglio formare*/

```
/*metodo non gerarchico kmedie*/
cluster kmeans v1-v6, k(2) name(cluster2)   /*2 cluster*/
cluster kmeans v1-v6, k(3) name(cluster3)   /*3 cluster*/
cluster kmeans v1-v6, k(4) name(cluster4)   /*4 cluster*/
```

/*per decidere tra le opzioni 2, 3, 4 cluster calcolo l'indicatore Calinski Harabasz pseudo-F, ancora una volta mi fermo quando l'indicatore è più alto*/

```
cluster stop cluster2
cluster stop cluster3
cluster stop cluster4
```

Number of clusters	Calinski/ Harabasz pseudo-F
2	13.76

```
. cluster stop cluster3
```

Number of clusters	Calinski/ Harabasz pseudo-F
3	26.56

```
. cluster stop cluster4
```

Number of clusters	Calinski/ Harabasz pseudo-F
4	21.84

/*se avessi una variabile quantitativa, per esempio il reddito degli individui potrei fare
anova reddito cluster 3
per vedere se i gruppi sono significativamente diversi anche rispetto al reddito*/