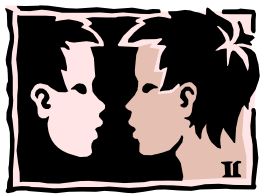
Analisi multivariata
a.a. 2011/12

Analisi dei gruppi (Cluster Analysis)

Carla Rampichini
Dipartimento di Statistica 'G. Parenti'

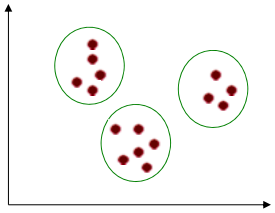
Obiettivo principale

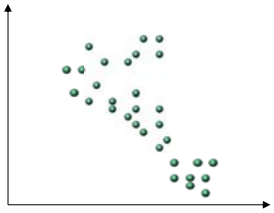
Considerare un gran numero di unità statistiche e creare un certo numero di gruppi distinti che contengono unità simili, in base a tutte le variabili considerate



Esempi di cluster

- Gruppi ben definiti
- Gruppi non ben definiti





Suggerimenti…

- Scatter-plot delle osservazioni nel caso 2 o 3 variabili
- Scatter-plot delle prime 2 o 3 componenti principali

Fasi del processo di analisi dei gruppi

1. Scelta delle unità di osservazione;
2. Scelta delle variabili; — Operazioni preliminari
3. Omogeneizzazione scale di misura;
4. Scelta della **misura di similarità** o diversità tra unità statistiche;
5. numero di gruppi; — Costruzione dei gruppi
6. Scelta del **criterio di raggruppamento**;
7. Scelta dell'**algoritmo di classificazione** ;
8. Interpretazione dei risultati ottenuti.

Scelta delle unità

- Popolazione completa ➔analisi descrittiva;
- Campione ➔inferenza

Ponderazione eventuale delle unità statistiche (più importante nel caso di dati campionari)

Selezione delle variabili

Criterio di scelta

Le variabili selezionate dovrebbero descrivere la somiglianza tra unità statistiche relativamente al problema di ricerca affrontato

Intuizione e giudizio del ricercatore
Importantissimi !!!

Considerare:
- Ricerche già fatte
- Teoria
- Ipotesi da verificare

Ponderazione delle variabili

- Esplicite: definite a priori per assegnare maggiore importanza ad alcune variabili;
- Implicite:
  - Varianze diverse;
  - Correlazioni
  - ...

Misure di similarità o distanza

Quanto sono *vicine* o *lontane* le u.s.?

- Distanza tra due u.s.: osservazioni a distanza minore sono più simili!
- Esistono molte distanze, la scelta non è neutrale: la più usata è la distanza euclidea;
- Problemi di standardizzazione e di ponderazione implicita.

Indice di distanza

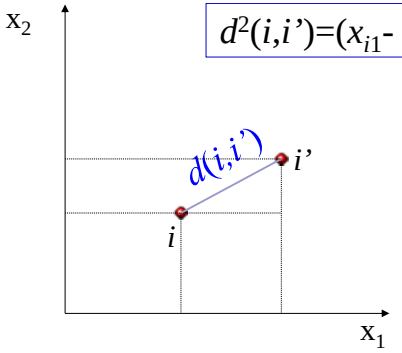
$d(i,i'): \Omega \rightarrow \mathbb{R}^+, \quad \Omega = \{(i,i'), i,i'=1,2,\ldots,n\}$

Proprietà

1. $d(i,i') >= 0$ 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''

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

Ponderazione implicita

L'importanza di ogni variabile nella determinazione della distanza è proporzionale alla varianza

Esempio

Matrice dei dati

45	30000
43	35000
47	34000

matrice delle distanze euclidee

0	5000	4000
	0	1000
		0

$d(1,2)=((45-43)^2+(30000-35000)^2)^{1/2}=5000$

Vettore delle medie=(45 33000)' varianze: V ($\mathbf{x}_1$)=2.6, V ($\mathbf{x}_2$)=4666666

Importanza di $\mathbf{x}_1$ sulle distanze è trascurabile!!

Distanza euclidea ponderata (➔cambiamento unità di misura)

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

pesi

$$d_w(i,i')^2 = (\mathbf{x}_i - \mathbf{x}_{i'})' \mathbf{D}_w (\mathbf{x}_i - \mathbf{x}_{i'})$$

$\mathbf{D}_w$=diag(w_s) matrice dei pesi

Se $w_j=1/\sigma_j^2$➔addendi di d_w sono numeri puri

segue esempio ponderazione implicita…

Per assegnare lo stesso peso alle variabili si può procedere alla standardizzazione

$z=(x-\mu)/\sigma$, $d(\mathbf{z}_i, \mathbf{z}_{i'})=d_w(\mathbf{x}_i, \mathbf{x}_{i'})$

Esempio

dati standardizzati

0.00	-1.39
-1.22	0.93
1.22	0.46

matrice delle distanze euclidee

0	2.62	2.22
	0	2.48
		0

L'ordinamento tra le distanze è cambiato!!

Standardizzazione elimina unità di misura, ma *diluisce* le differenze tra gruppi rispetto alle variabili più discriminanti.

Distanza euclidea ponderata

Sensibile alla "dimensione" delle u.s. per l'omogeneità: $d(\lambda\mathbf{x}_i, \lambda\mathbf{x}_{i'})=\lambda d(\mathbf{x}_i, \mathbf{x}_{i'})$

Esempio

Classificazione crani (adulti e bimbi) homo-sapiens *vs* gorilla. La distanza euclidea mette crani dei bimbi (λ più piccolo) insieme: più vicini dei crani appartenenti alla stessa specie.

la dimensione delle unità statistiche può oscurare altre differenze!!!

Metrica di Minkowski

$$d_m(i,i') = \left[\sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right|^m \right]^{1/m}$$

- m=1 ➔ distanza 'city-block'
- m=2 ➔ distanza euclidea
- Al variare di m cambia il peso attribuito a differenze grandi e piccole

Distanze per variabili non negative

- Distanza di Canberra

$$d_c(i,i') = \sum_{j=1}^{p} \frac{\left| x_{ij} - x_{i'j} \right|}{\left(x_{ij} + x_{i'j} \right)}$$

- Distanza di Czekanowski

$$d_{Cz}(i,i') = 1 - \frac{2\sum_{j=1}^{p} \min(x_{ij}, x_{i'j})}{\sum_{j=1}^{p} (x_{ij} + x_{i'j})}$$

Distanza di Mahalanobis

$$d_M(i,i') = \left[\left(\mathbf{x}_i - \mathbf{x}_{i'} \right)' \mathbf{S}^{-1} \left(\mathbf{x}_i - \mathbf{x}_{i'} \right) \right]^{1/2}$$

- Tiene conto della correlazione tra variabili
- Se le correlazioni sono nulle coincide con la distanza euclidea calcolata sulle variabili standardizzate
- Per p=2, il luogo dei punti equidistanti dal centroide secondo la distanza di Mahalanobis è un'ellissi, tanto più allungata quanto più |r| è grande
- È invariante per trasformazioni di scala
- Più complessa da calcolare e di interpretazione meno immediata rispetto alla distanza euclidea

Correlazione

- Se c'è correlazione tra le variabili, alcune possono avere una ponderazione implicita superiore ad altre, nonostante la standardizzazione.
- Distanza di Mahalanobis tiene conto delle correlazioni :

$$d_w(i,i')^2 = (\mathbf{x}_i - \mathbf{x}_{i'})' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_{i'})$$

coincide con distanza euclidea su variabili standardizzate in caso di incorrelazione, ma attenua differenze tra gruppi.

Metrica

- Un indice di distanza si dice una metrica se soddisfa la disuguaglianza triangolare

date tre unità qualsiasi i i' e i'' risulta sempre
$d(i,i') \leq d(i',i'') + d(i'',i)$

- cioè la distanza che intercorre tra due punti è sempre minore della somma delle distanze tra tali punti e un terzo punto.
- Questa proprietà, naturalissima nella nostra percezione delle distanze spaziali, non è sempre verificata per certi indici di distanza in spazi astratti.
- La distanza Euclidea è una metrica

Ultrametrica

- Una distanza $d(i,i')$ si dice distanza ultrametrica se gode della seguente proprietà, detta *disuguaglianza ultrametrica*
  date tre unità qualsiasi *i, i'* e *i"* si ha

$$d(i,i') \leq \max \{d(i',i''), d(i'',i)\}$$

- La disuguaglianza ultrametrica richiede che la massima distanza tra l'unità *i"* e la coppia di unità (i,i') non possa mai scendere al di sotto della distanza che separa *i* e *i'*.

- Si osservi che se $d(i,i')$ è una distanza ultrametrica allora è automaticamente una metrica, perchè la disuguaglianza ultrametrica implica la disuguaglianza triangolare.

La procedura DISTANCE

Esempi

- distanze_es8_1_zani.sas
- distanze_alimentari.sas

Indici di similarità per variabili dicotomiche

- Due u.s. possono essere confrontante in base alla presenza/assenza di una certa caratteristica
  ➔considerare tante dummy quante sono le caratteristiche

$$x_{ij} = \begin{cases} 1 & \text{se } i \text{ posside la caratteristica } j \\ 0 & \textit{altrimenti} \end{cases}$$

- La distanza euclidea in questo caso 'conta' il n. di variabili discordanti

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

- NB: le coppie 0-0 hanno lo stesso peso delle coppie 1-1

Indice di Gower

indice di somiglianza generale, valido per dati quantitativi e qualitativi

$$s(i,i')=\Sigma_j\, c_{ii'j}/\Sigma_j\, w_{ii'j}$$

- $c_{ii'j}$ è una misura di somiglianza tra i e i' tenuto conto solo del carattere j, j=1,…,p
- $w_{ii'j}$ è un peso che può assumere solo valori 1 e 0, vale 0 quando non è sensato un confronto tra i e i' per quel carattere e 1 altrimenti.

Indice di Gower (2)

Se X_j è un carattere quantitativo:

$$c_{ii'j} = 1 - |x_{ij} - x_{i'j}|/R_j$$

dove R_j è il campo di variazione della variabile j

Se X_j e' un carattere dicotomico:

Unità *i*	1 1 0 0
Unità *i'*	1 0 1 0
$c_{ii'j}$	1 0 0 0
$w_{ii'j}$	1 1 1 0

➔dai confronti vengono esclusi i casi in cui il carattere è assente (0) in entrambe le unità, mentre la somiglianza $c_{ii'j}$ vale 1 se vi è co-presenza del carattere.

- Se tutti i caratteri sono dicotomici *s(i; i')* coincide l'indice di somiglianza di Jaccard.

Indici di similarità per caratteri dicotomici

- Sono state fatte molte proposte in letteratura per pesare diversamente le coppie 0-0 e 1-1 (due persone che sanno suonare il pianoforte evidenziano una somiglianza maggiore tra due unità rispetto a due persone che NON sanno suonare il piano!)

$$S(i,i') = \frac{a}{p}$$ Indice di Russel e Rao

		u.s. i'		
		1	0	tot
u.s. i	1	a	b	a+b
	0	c	d	c+d
	tot	a+c	b+d	p

$$S(i,i') = \frac{a}{a+b+c}$$ I. di Jaccard

$$S(i,i') = \frac{a+d}{p}$$ I. di Sokal e Michener

esempio in distanze.sas

Indice di Gower (3)

Se X_j è un carattere politomico:

Unità *i*	a b 0 0
Unità *i'*	a a b 0
$c_{ii'j}$	1 0 0 1
$w_{ii'j}$	1 1 1 1

$w_{ii'j}$ =1 sempre (salvo in caso di dato mancante)

$c_{ii'j}$ = 1 se le due unità hanno la stessa modalità del carattere, e zero altrimenti.

Se questa definizione viene applicata a dati dicotomici si ottiene l'indice di Sokal e Michener per il quale $c_{ii'j}$ = 1 anche nella situazione di co-assenza del carattere nelle due unità.

Vantaggi indice di Gower

- Si può dimostrare che la distanza definita da

$$d(i,i') = 2\sqrt{1 - s(i,i')}$$

e' una metrica (cioè soddisfa la disuguaglianza triangolare) ed esiste una configurazione di punti per i quali essa è una distanza Euclidea

- Inoltre Gower ha dimostrato che la matrice delle somiglianze di elemento generico s(i; i') è semidefinita positiva e questa proprietà è fondamentale per utilizzare i metodi di scaling multidimensionale

Strutture di classificazione

- Una volta definito un indice di prossimità è necessario introdurre una definizione precisa del concetto di gruppo.
- Il miglior modo per farlo è quello di stabilire delle strutture matematiche tali da poter essere utilizzate per la classificazione.
- Le due strutture più comunemente utilizzate sono le partizioni e le gerarchie.

Partizioni

- Per X tutte quantitative, u_i è un punto, $\mathbf{x}_i$, nello spazio euclideo;
- Partizione generata da G punti: $m_1,\ldots,m_G$

$$A_{g^*}=\{\mathbf{x}_i : d(\mathbf{x}_i, m_{g^*})=\min d(\mathbf{x}_i, m_g),\ g=1,\ldots,G\}$$

In ogni classe A_{g^*} si trovano le u.s. che sono più vicine a m_{g^*} che agli altri punti che generano la partizione

Gerarchie

- Vedere dispense Marchetti e il file metodi_gerarchici.pdf
- Gerarchie
- Alberi gerarchici
- Ultrametrica associata e dendrogramma

Come formare i gruppi?

- Definire un indice di distanza
- Sintetizzare l'informazione contenuta nella matrice delle distanze attraverso indici che misurino l'omogeneità e la separazione delle classi (delle partizioni o delle gerarchie).
- Calcolare gli indici per *tutte* le partizioni possibili o tutte le gerarchie possibili delle *n* unità: però il numero di partizioni o di gerarchie da considerare è troppo grande, anche per valori piccoli di *n* (p.e. per n=20 sono possibili $5{,}17*10^{13}$ partizioni e $5{,}64*10^{29}$ gerarchie!)

- Restringere la ricerca ad un sottoinsieme (molto ridotto) delle partizioni o delle gerarchie.

Metodi di formazione dei gruppi

Ce ne sono moltissimi…

- Metodi gerarchici:
  - di tipo agglomerativo
  - di tipo disgiuntivo
- Metodi non gerarchici:
  - Partizioni
  - Classi sovrapposte

Indice di aggregazione

- Dati due gruppi A,B si definisce $D(A,B)$ indice di distanza tra gruppi
- $D(A,B)$ è una funzione reale positiva tale che:
  - $D(A,B) = D(B,A)$
  - misura la distanza tra i due gruppi sulla base delle distanze tra unità.

Algoritmo gerarchico agglomerativo

- Si costruisce la matrice delle distanze tra le unità
- Si definisce un'*indice di aggregazione* $D(A,B)$

1. Si parte dalla partizione banale P_0 con gruppi di un solo elemento.
2. Si costruisce una nuova partizione unendo i due gruppi della partizione precedente che minimizzano l'indice di aggregazione $D(A,B)$
3. Si ripete il passo 2 fino a riunire tutti i gruppi in uno solo.

Indici di aggregazione monotoni

- Al passo $t-1$ dell'algoritmo, le due classi $A(t-1)$ e $B(t-1)$ che minimizzano $D(A,B)$ vengono fuse in una sola, $C(t)$ e formano un nodo del dendrogramma con valore della graduazione

$$h(C(t)) = h_t = D(A(t-1);B(t-1))$$

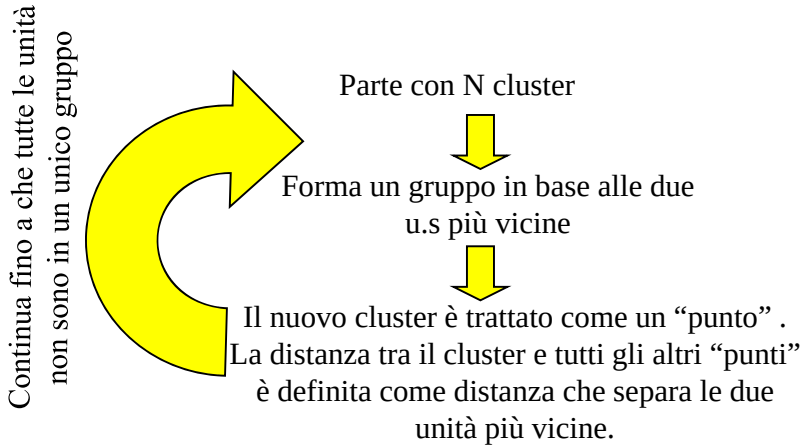
Poiché $t = 0, \ldots, n-1$ ➔ si ottengono n valori $h_0; h_1, \ldots, h_{n-1}$

l'indice di aggregazione è monotono se risulta

$$0 = h_0 \leq h_1 \leq \ldots \leq h_{n-1}$$

- In caso contrario per qualche valore di t avviene un inversione, cioè l'indice per i due gruppi che si fondono $h(A \cup B)$ è minore dell'indice per i due gruppi che si sono fusi un passo precedente. Gli indici NON monotoni sono scarsamente utilizzati perché poco interpretabili.

Metodo del legame singolo



Esempio metodo del legame singolo

Matrice delle distanze tra sei punti

	1	2	3	4	5	6
1		0.31	0.23	0.32	0.26	0.25
2			0.34	0.21	0.36	0.28
3				0.31	0.04	0.07
4					0.31	0.28
5						0.09

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

$$D(A,B) = \min_{i \in A, i' \in B} \{d(i,i')\}$$

cluster11.sas

esempio metodo legame singolo (2)

Matrice delle distanze tra 5 punti

	1	2	[3,5]	4	6
1		0.31	0.23	0.32	0.25
2			0.34	0.21	0.28
[3,5]				0.31	0.07
4					0.28
6					

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

Distanza più piccola

$$D([3,5],2) = \min_{i \in [3,5],2} \{d(3,2), d(5,2)\} = 0.34$$

esempio metodo legame singolo (3)

Matrice delle distanze tra quattro punti

	1	2	[3,5,6]	4
1		0.31	0.23	0.32
2			0.28	0.21
[3,5,6]				0.28
4				

Distanza più piccola

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

C_3={[1],[2,4],[3,5,6]}

esempio metodo legame singolo (4)

Matrice delle distanze tra tre "punti"

	1	[2,4]	[3,5,6]
1		0.31	0.23
[2,4]			0.28
[3,5,6]			

Distanza più piccola

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

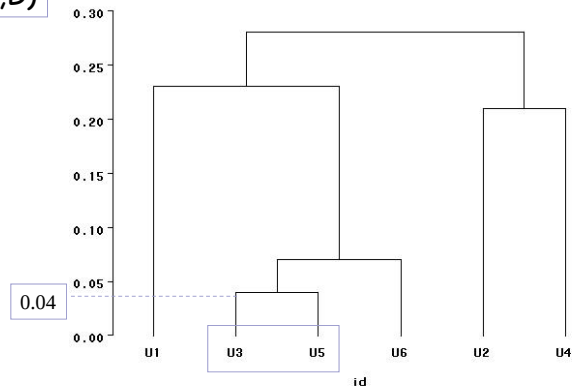
C_3={[1],[2,4],[3,5,6]}

C_4={[2,4],[1,3,5,6]}

Un unico gruppo

Dendrogramma

h(.)=D(A,B)



Costruzione della gerarchia (output SAS cluster11.sas)

```
La procedura CLUSTER
Analisi dei cluster con legame singolo

                Cronologia dei cluster

                                                          T
                                                Dist      i
 NCL     ---Cluster uniti----       Freq        min      e

   5     U3           U5                2      0.04
   4     CL5          U6                3      0.07
   3     U2           U4                2      0.21
   2     U1           CL4               4      0.23
   1     CL2          CL3               6      0.28
```

Esempio: alcuni indicatori per 50 stati americani (Marchetti, 1977)

The MEANS Procedure

Variable	Label	Mean	Coeff of Variation
pop75	popolazione 1/7/75 (migliaia)	4246.42	105.14
reddito	reddito pro-capite 1974	4435.80	13.85
analfab	%analfabeti sulla popolazione	1.17	52.10
vita	vita media in anni	70.88	1.89
crimini	% crimini per 100000 ab.	7.38	50.03
diploma	% diplomati 1970	53.11	15.21
area	area dello stato (miglia quadr.)	70735.88	120.63
freddo	temperatura media minima (F°)	104.46	49.76

cluster.sas

Matrice dei dati

Obs	stato	pop75	reddito	analfab	vita	crimini	diploma	area	freddo
1	alabama	3615	3624	2.10	69.05	15.10	41.30	50708	20
2	alaska	365	6315	1.50	69.31	11.30	66.70	566432	152
3	arizona	2212	4530	1.80	70.55	7.80	58.10	113417	15
4	arkansas	2110	3378	1.90	70.66	10.10	39.90	51945	65
5	california	21198	5114	1.10	71.71	10.30	62.60	156361	20
6	colorado	2541	4884	0.70	72.06	6.80	63.90	103766	166
7	connecticut	3100	5348	1.10	72.48	3.10	56.00	4862	139
8	delaware	579	4809	0.90	70.06	6.20	54.60	1982	103
9	florida	8277	4815	1.30	70.66	10.70	52.60	54090	11
10	georgia	4931	4091	2.00	68.54	13.90	40.60	58073	60
11	hawaii	868	4963	1.90	73.60	6.20	61.90	6425	0
12	idaho	813	4119	0.60	71.87	5.30	59.50	82677	126
13	illinois	11197	5107	0.90	70.14	10.30	52.60	55748	127
14	indiana	5313	4458	0.70	70.88	7.10	52.90	36097	122
15	iowa	2861	4628	0.50	72.56	2.30	59.00	55941	140
16	kansas	2280	4669	0.60	72.58	4.50	59.90	81787	114
17	kentucky	3387	3712	1.60	70.10	10.60	38.50	39650	95
18	lousiana	3806	3545	2.80	68.76	13.20	42.20	44930	12
19	maine	1058	3694	0.70	70.39	2.70	54.70	30920	161
20	maryland	4122	5299	0.90	70.22	8.50	52.30	9891	101
21	massachusset	5814	4755	1.10	71.83	3.30	58.50	7826	103
22	michigan	9111	4751	0.90	70.63	11.10	52.80	56817	125
23	minnesota	3921	4675	0.60	72.96	2.30	57.60	79289	160
24	mississipi	2341	3098	2.40	68.09	12.50	41.00	47296	50
25	missuori	4767	4254	0.80	70.69	9.30	48.80	68995	108
.....									
50	wyoming	376	4566	0.60	70.29	6.90	62.90	97203	173

Proprietà metodo legame singolo

- Indice di aggregazione monotono
- Metodo di classificazione ordinale
- VANTAGGI: permette di individuare gruppi di qualsiasi forma, purché ben separati.
- SVANTAGGI: *effetto di concatenamento*, cioè ad ogni fusione le unità non ancora classificate tendono ad essere incorporate in gruppi già esistenti piuttosto che formare nuovi gruppi.
- Due gruppi possono essere aggregati nei primi passi ed essere considerati poco dissimili anche solo perchè esiste una catena di unità che unisce i due gruppi.

Altri metodi gerarchici

- Metodo del legame completo
- Metodo del legame medio
- Metodo del centroide
- Metodo di WARD

Metodo del legame completo

Matrice delle distanze tra sei punti

	1	2	3	4	5	6
1		0.31	0.23	0.32	0.26	0.25
2			0.34	0.21	0.36	0.28
3				0.31	0.04	0.07
4					0.31	0.28
5						0.09
6						

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

$$D(A,B) = \max_{i \in A, i' \in B} \{d(i,i')\}$$

Metodo del legame completo

Matrice delle distanze tra 5 punti

	1	2	[3,5]	4	6
1		0.31	0.26	0.32	0.25
2			0.36	0.21	0.28
[3,5]				0.32	0.09
4					0.28
6					

Distanza più piccola

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

$$D([3,5],2) = \max_{i \in [3,5],2} \{d(3,2), d(5,2)\} = 0.36$$

Metodo del legame completo

Matrice delle distanze tra quattro punti

	1	2	[3,5,6]	4
1		0.31	0.26	0.32
2			0.36	0.21
[3,5,6]				0.32

Distanza più piccola

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

C_3={[1],[2,4],[3,5,6]}

Metodo del legame completo

Matrice delle distanze tra tre “punti”

	1	[2,4]	[3,5,6]
1		0.32	0.23
[2,4]			0.36
[3,5,6]			

Distanza più piccola

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

C_3={[1],[2,4],[3,5,6]}

C_4={[2,4],[1,3,5,6]}

Un unico gruppo

Proprietà metodo del legame completo

- Indice di aggregazione monotono
- Metodo di classificazione ordinale

Metodo del legame medio

Matrice delle distanze tra sei punti

	1	2	3	4	5	6
1		0.31	0.23	0.32	0.26	0.25
2			0.34	0.21	0.36	0.28
3				0.31	0.04	0.07
4					0.31	0.28
5						0.09
6						

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

$$D(A,B) = \sum_{i \in A} \sum_{i' \in B} d(i,i') / n_A n_B$$

Metodo del legame medio

Matrice delle distanze tra 5 punti

	1	2	[3,5]	4	6
1		0.31	0.245	0.32	0.25
2			0.35	0.21	0.28
[3,5]				0.31	0.08
4					0.28
6					

Distanza più piccola

C_0={[1],[2],[3],[4],[5],[6]}

C_1={[1],[2],[3,5],[4],[6]}

C_2={[1],[2],[3,5,6],[4]}

$$D([3,5],1) = \left[d(3,2) + d(5,2) \right] / 2 = 0.245$$

Proprietà metodo del legame medio

- Indice di aggregazione monotono
- Metodo di classificazione NON ordinale: utilizza valore delle distanze per calcolare indice di aggregazione

Metodo di Ward

$$D(A,B) = \frac{N_A N_B}{N_A + N_B} d^2(\bar{\mathbf{x}}_A, \bar{\mathbf{x}}_B)$$

Vettori delle medie dei gruppi A e B

L'indice di Ward:

- misura la parte della dispersione di A∪B dovuta alle differenze tra i gruppi.
- è monotono
- il metodo che ne deriva è NON ordinale.

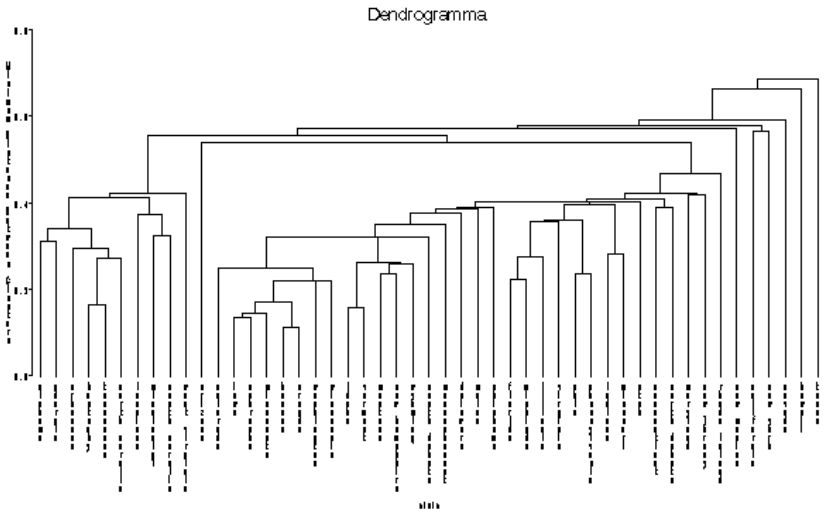
Esempio: alcuni indicatori per 50 stati americani (Marchetti, 1977)

The MEANS Procedure

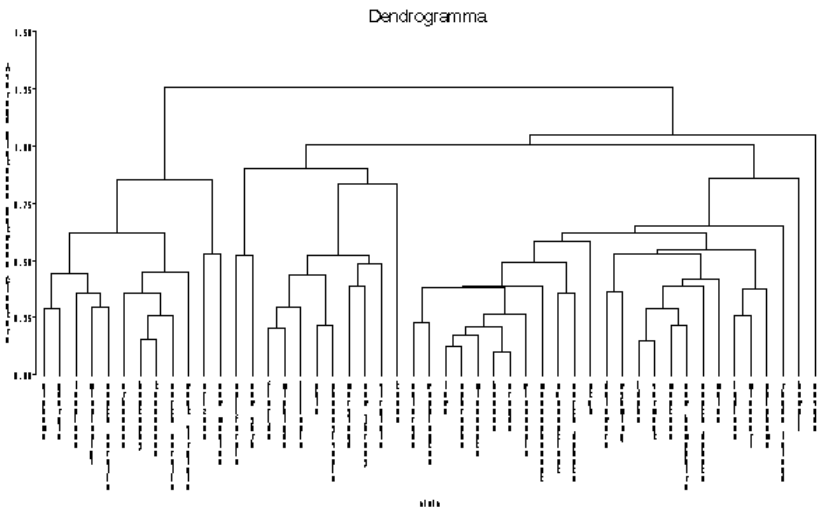
Variable	Label	Mean	Coeff of Variation
pop75	popolazione 1/7/75 (migliaia)	4246.42	105.14
reddito	reddito pro-capite 1974	4435.80	13.85
analfab	%analfabeti sulla popolazione	1.17	52.10
vita	vita media in anni	70.88	1.89
crimini	% crimini per 100000 ab.	7.38	50.03
diploma	% diplomati 1970	53.11	15.21
area	area dello stato (miglia quadr.)	70735.88	120.63
freddo	temperatura media minima (F°)	104.46	49.76

cluster.sas

Metodo del legame semplice



Metodo del legame medio

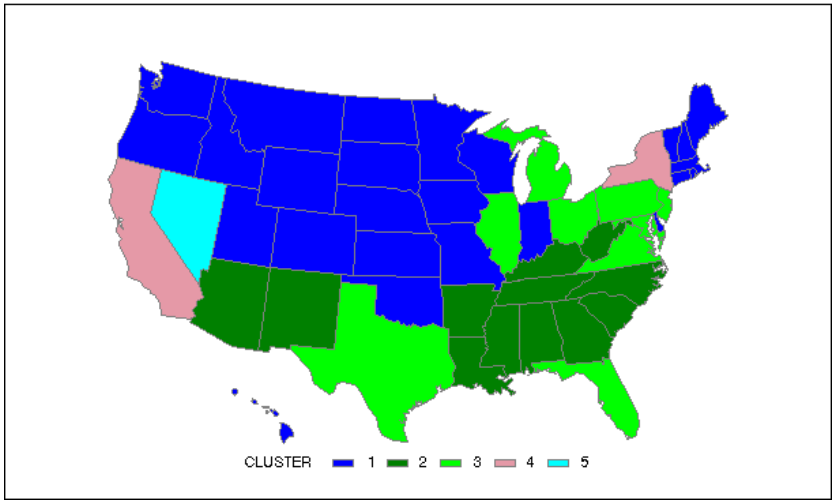


Caratterizzazione dei gruppi

The MEANS Procedure

CLUSTER	N Obs	Variable	Mean	Std Dev
1	25	pop75	2178.56	1674.02
		reddito	4506.96	393.38
		analfab	0.78	0.31
		vita	71.83	0.94
		crimini	4.42	2.01
		diploma	57.21	4.99
		area	56500.28	38944.68
		freddo	127.20	46.94
2	12	pop75	3147.92	1305.46
		reddito	3710.58	357.09
		analfab	2.00	0.39
		vita	69.40	0.95
		crimini	11.11	2.41
		diploma	43.04	6.54
		area	55987.67	30212.62
		freddo	62.67	34.27

Mappa dei gruppi

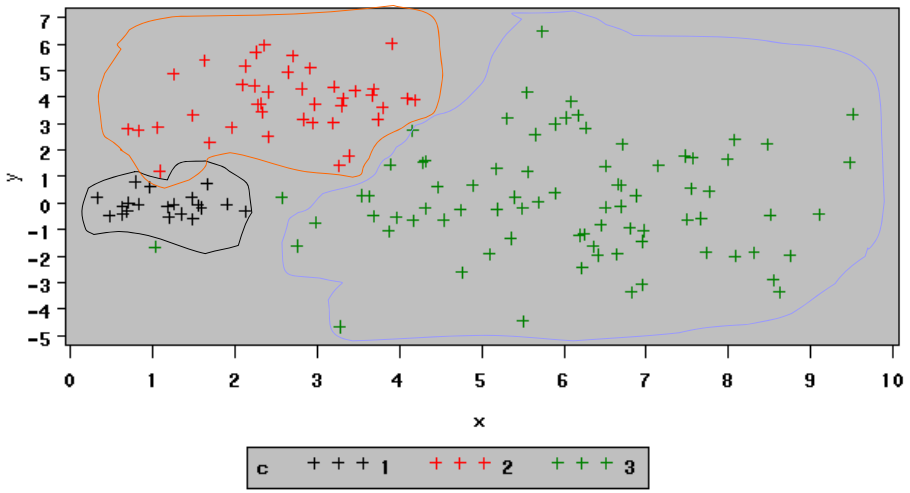


Quale metodo scegliere?

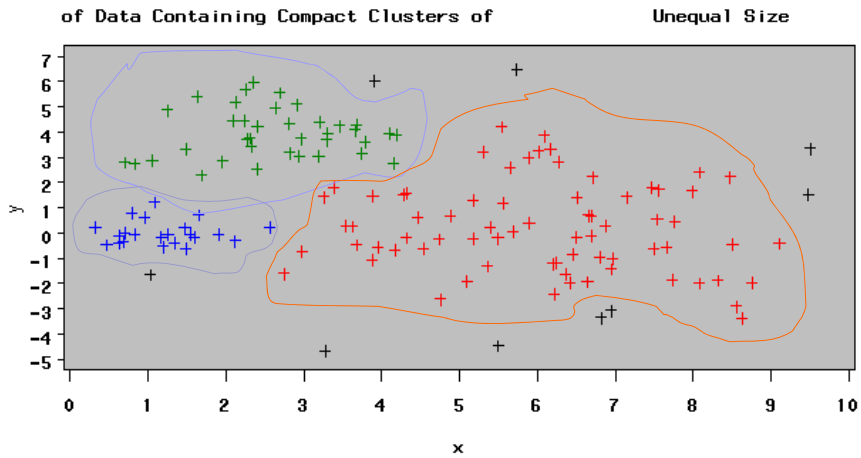
Alcuni esempi di come le varie tecniche individuano i gruppi sono presentati attraverso dati simulati:

- Gruppi compatti e ben identificabili `cluster4.sas`
- Gruppi non ben definiti `cluster5.sas`, `cluster6.sas`
- Gruppi diversa dimensione `cluster7.sas`, `cluster8.sas`
- Gruppi allungati `cluster9.sas`

Gruppi veri da dati Multinormali: gruppi di dimensione diversa

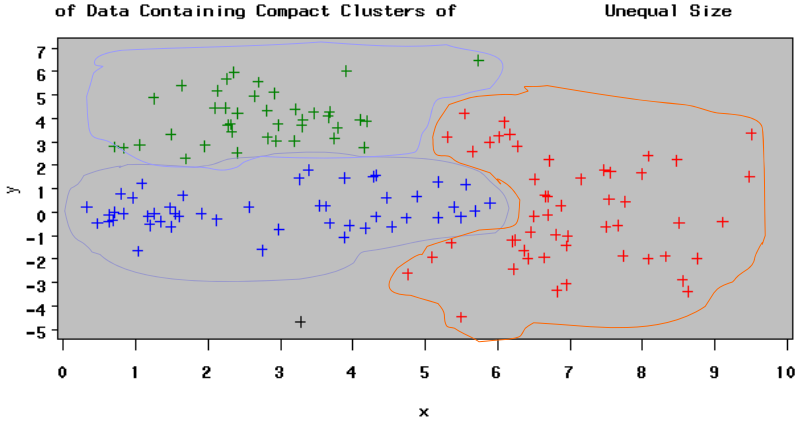


Single Linkage Cluster Analysis



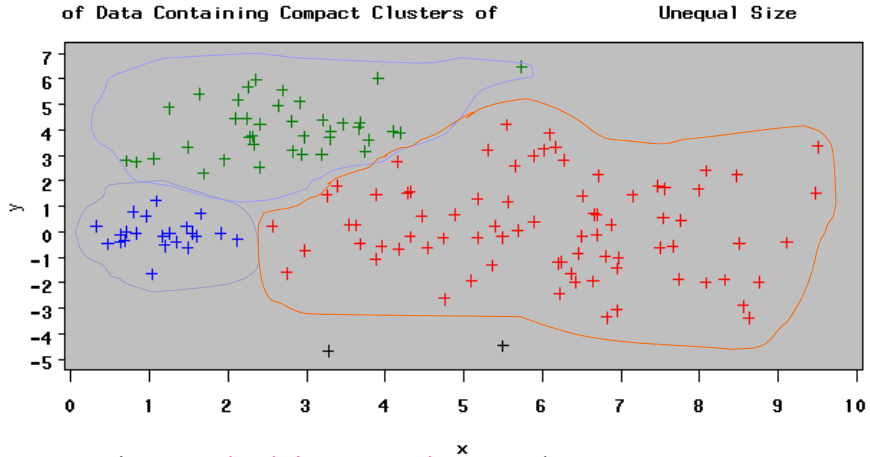
Il metodo del legame singolo riproduce correttamente i gruppi 'veri'.

Average Linkage Cluster Analysis



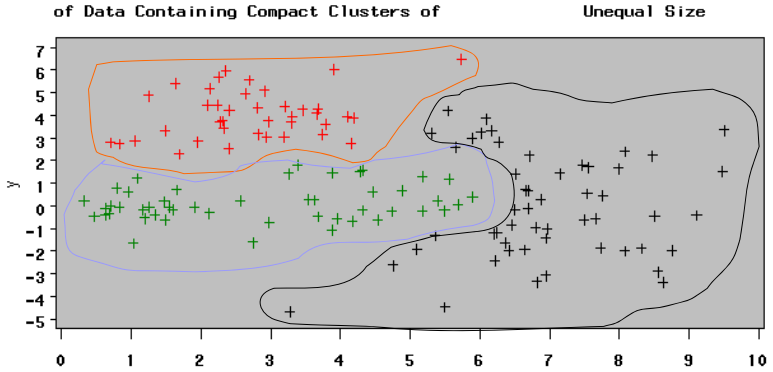
Con il metodo del legame medio i due gruppi di sinistra 'entrano' nel gruppo di destra, rendendo le varianze di ciascun gruppo più simili tra loro, rispetto a quelle dei gruppi'veri'.

Centroid Cluster Analysis

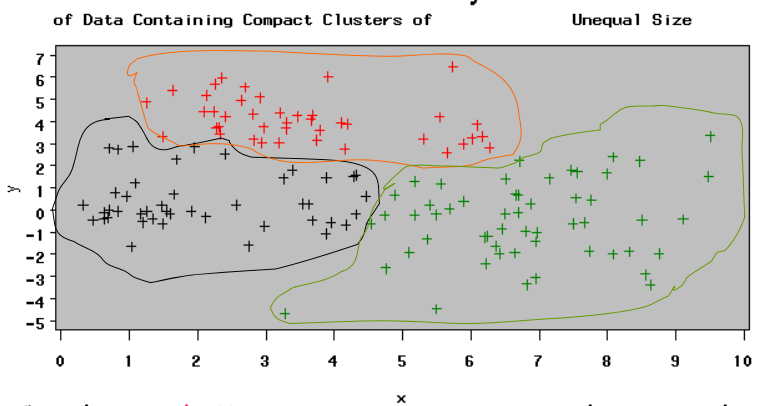


Il metodo del centroide riproduce abbastanza bene i gruppi 'veri'.

Ward's Minimum Variance Cluster Analysis

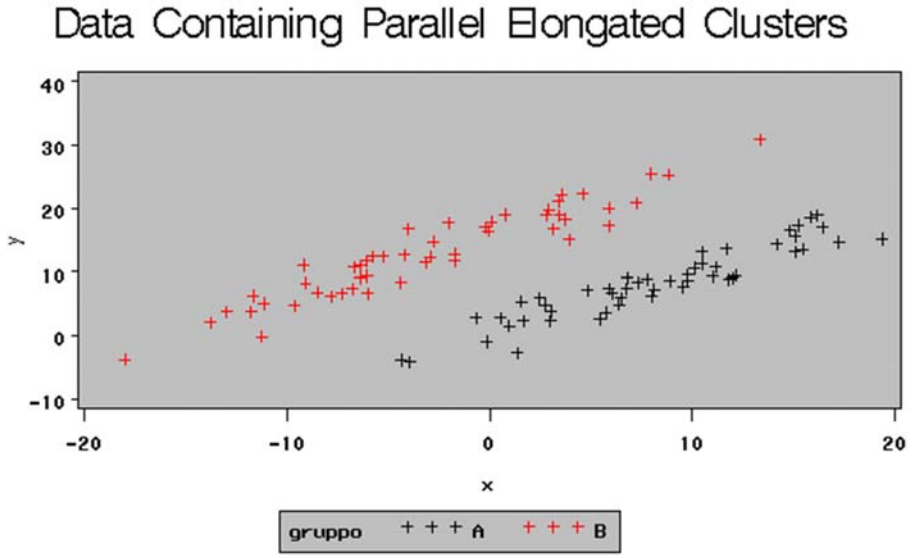


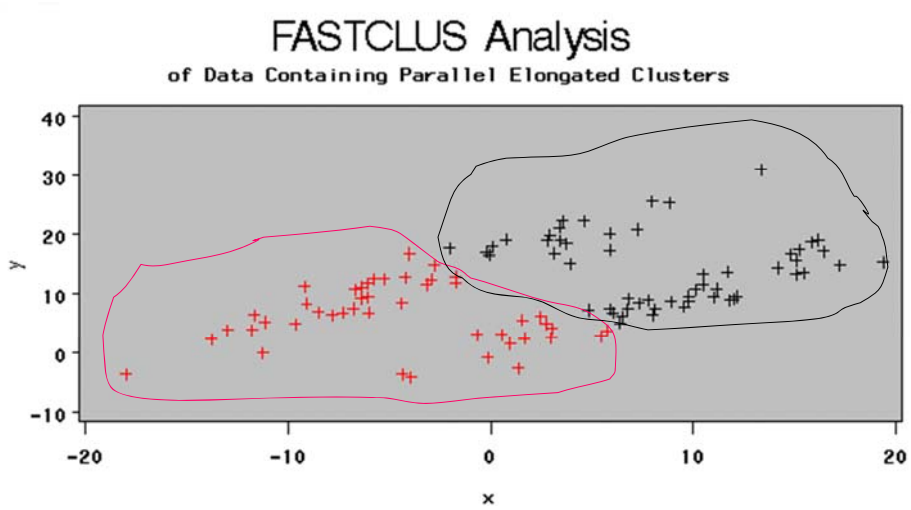
Con il metodo di Ward il gruppo in alto a sin è separato correttamente, ma il gruppo in basso a sin comprende buona parte delle osservazioni del gruppo di destra: questo errore è dovuto alla tendenza del metodo a creare gruppi di uguale dimensione.



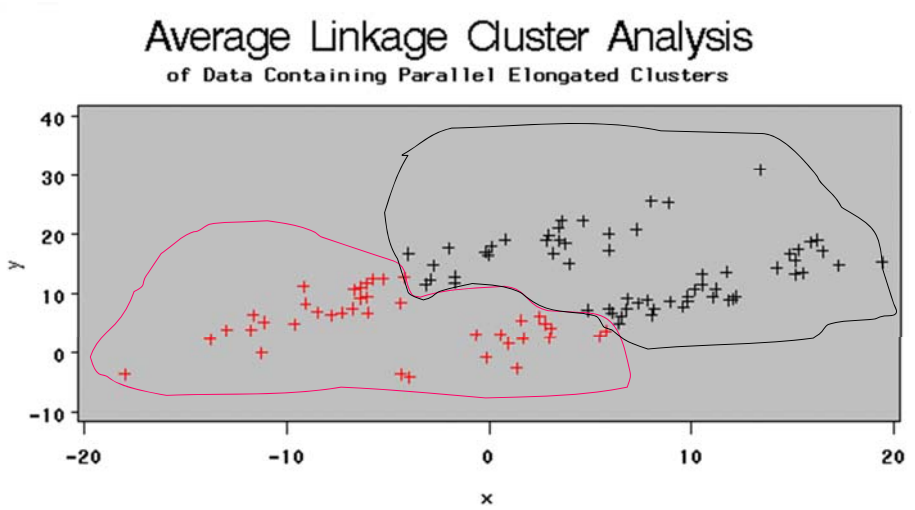
Con il metodo K-means si ottiene un risultato simile a quello ottenuto con il metodo di Ward: questo errore è dovuto alla tendenza del metodo a creare gruppi di uguale dimensione.

Gruppi veri allungati

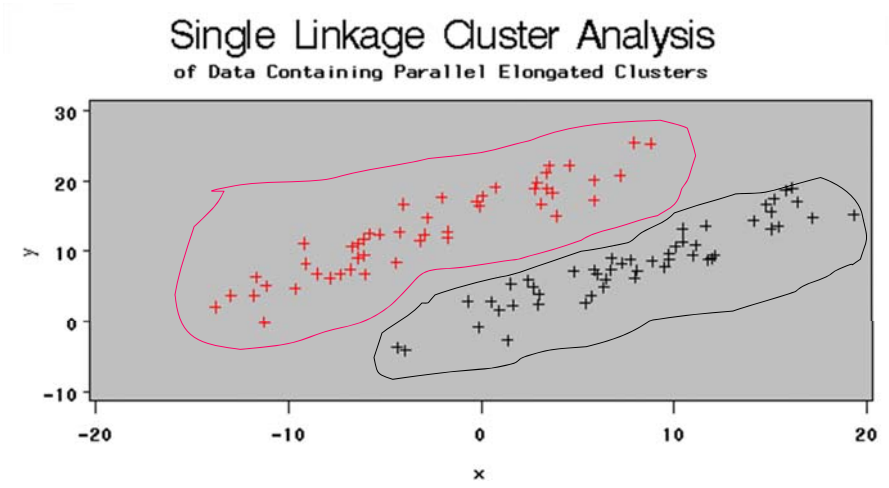




Il metodo K-means forma due gruppi sferici!!



Anche il metodo del legame medio forma due gruppi sferici!! (analogo risultato si ottiene con i metodi di Ward e del centroide)



Il metodo del legame singolo aggrega correttamente le unià nei due gruppi 'veri'.

Cluster analysis con STATA

- CLUSTER analisi dei gruppi
- CLUSTERMAT analisi matrice di dissimilarità

- Vari metodi gerarchici e non gerarchici disponibili
- Varie misure di distanza e dissimilarità possibili
- Esempi: **hierarchical.do**, **kmeans_cluster.do**

Procedure SAS per l'analisi dei gruppi

- **CLUSTER**
  analisi gerarchica dei gruppi (12 metodi)
- **FASTCLUS**
  partizione individuata con il metodo k-means
- **MODECLUS**
  partizione con densità non parametrica
- **VARCLUS**
  analisi dei gruppi sulle variabili
- **TREE**
  costruisce il dendrogramma e individua i gruppi

Procedura CLUSTER

PROC CLUSTER METHOD = name < options > **;**
BY variables **;**
COPY variables **;**
FREQ variable **;**
ID variable **;**
RMSSTD variable **;**
VAR variables **;**

Scelta del metodo METHOD=

AVERAGE | AVE	legame medio
CENTROID | CEN	centroide
COMPLETE | COM	legame completo
DENSITY | DEN	densità non parametrica
EMLFLEXIBLE | FLE	ML gerarchico
MCQUITTY | MCQ	analisi di similarità
MEDIAN | MED	metodo di Gower
SINGLE | SIN	legame singolo
TWOSTAGE | TWO	densità a due stadi
WARD | WAR	metodo di Ward

Opzioni di PROC CLUSTER (selezione)

STANDARD | STD standardizza le variabili

OUTTREE=*SAS-data-set*
crea data set per PROC TREE

SIMPLE | S calcola indici univariati

DATA=*SAS-data-set*
coordinate o TYPE=DISTANCE

Esempio forni a microonde: **cluster_microonde.sas**

Dendrogramma

PROC TREE < options > **;**
NAME variables **;**
HEIGHT variable **;**
PARENT variables **;**
BY variables **;**
COPY variables **;**
FREQ variable **;**
ID variable **;**

Opzioni di PROC TREE (selezione)

NCLUSTERS=*n* n. gruppi per OUT=
OUT=*SAS-data-set*
HORIZONTAL diagramma orizzontale
HORDISPLAY=RIGHT foglie a dx

+ molte altre opzioni grafiche!

Dendrogramma forni a microonde

- confronto dendrogramma ottenuto con 3 metodi
  - Ci sono due gruppi molto omogenei al loro interno e ben distinti che vengono individuati da tutti e 3 i metodi: {electrolux,panasonix} e {de longhi, samsung, moulinex}
  - Ocean viene aggregato a gruppi diversi a seconda del metodo utilizzato
  - Sharp e Candy sono molto diversi da tutti gli altri

Come scegliere la partizione (quanti gruppi)?

- Se il 'gradino' tra un passo e il successivo è molto alto ➔ gli elementi da 'fondere' sono molto distanti➔fermarsi PRIMA di questo 'salto'
- Nell'esempio: si 'taglia' l'albero prima del valore 0.5 (distanze normalizzate per metodo legame singolo e completo) ottenendo 5 gruppi

Metodo del centroide

- La distanza tra due gruppi è definita come

$$D(A,B) = d(\overline{\mathbf{x}}_A, \overline{\mathbf{x}}_B)$$

- Il centroide del nuovo gruppo che si forma è

$$centroide(A \cup B) = \frac{n_A \overline{\mathbf{x}}_A + n_B \overline{\mathbf{x}}_B}{n_A + n_B}$$

microonde_ward.sas

Output SAS per metodo di Ward

- SPRSQ (Semipartial R-Squared) misura la riduzione nella proporzione di varianza spiegata che si ha unendo due cluster. Corrisponde alla devianza tra gruppi divisa per la devianza totale
- RSQ coefficiente di correlazione multipla R^2 = dev tra/dev tot cioè proporzione di varianza spiegata dal raggruppamento
- BSS (Between-Cluster Sum of Squares) devianza between per i due cluster uniti

Output SAS
metodi di Ward e del centroide

- Si individuano come con gli altri metodi i gruppi:

{electrolux,panasonix} {de longhi,samsung,moulinex}

➔pattern *naturale* nell'insieme di dati

- Sharp e Candy si confermano outliers

Come scegliere la partizione?

- Ridotta quota di devianza 'entro' gruppi rispetto alla devianza 'tra' gruppi➔ R^2 = dev tra/dev tot più alto possibile (trade-off dimensione gruppi e R^2)
- RMSSTD (Root Mean Square Standard Deviation)

$$RMSSTD = \sqrt{\frac{W_h}{p(n_h - 1)}}, \quad W_h = \sum_{s=1}^{p}\sum_{i=1}^{n_h}(x_{is} - \bar{x}_{sh})^2$$

- Un forte incremento di RMSSTD segnala che si sono 'fusi' due gruppi molto eterogeni

Partizione ben strutturata minimale

- Per ottenere gruppi 'oggettivi' o 'naturali' si richiede che la massima distanza tra unità all'interno dei gruppi sia più piccola della minima distanza tra gruppi
- Una partizione P={ $C_1, C_2, \ldots, C_g$} di un insieme di n elementi $\mathcal{U}$={ $u_1, u_2, \ldots, u_n$} per i quali si è definita una distanza d, si dice *ben strutturata* se

$$\max\{d(i,j)\} < \min\{d(r,s)\}$$

per ogni coppia di unità (i,j) appartenenti allo stesso gruppo e (r,s) appartenenti a gruppi diversi
- Si dice partizione ben strutturata minimale la partizione ben strutturata con il minor numero di gruppi

Partizione ben strutturata minimale

- Per ogni matrice delle distanze esiste una e una sola partizione ben strutturata minimale (Castagnoli, 1978)
- I metodi del legame singolo, del legame completo e del legame medio ad un certo passo della classificazione gerarchica individuano la partizione ben strutturata minimale
- Nelle applicazioni la partizione ben strutturata minimale è spesso costituita da un numero eccessivo di gruppi

Altre proprietà dei metodi gerarchici

- Il metodo del legame singolo e del legame completo sono invarianti per trasformazione monotona crescente delle distanze, cioè forniscono la stessa successione di partizioni per ogni trasformazione monotona crescente delle distanze
- Il metodo del legame medio non è invariante!

Svantaggi dei metodi gerarchici

- Fonti di errore e variabilità non sono formalmente considerate➔metodi sensibili agli *outliers*
- Non c'è possibilità di *riallocare* un'unità che è stata classificata erroneamente in uno dei passi iniziali dell'algoritmo
- Conviene sempre provare più metodi e indici di distanza: se i risultati sono abbastanza simili allora si può pensare che esistano dei gruppi *naturali*
- Valori comuni nella matrice delle distanze (*ties*) possono portare a risultati diversi a seconda di come vengono trattati
- Il metodo del centroide può portare a *inversioni*

Metodi di raggruppamento NON gerarchici

- il numero di gruppi desiderato K deve essere noto prima o deve essere determinato come parte dell'algoritmo di classificazione
- Questi metodi sono utili per data set molto grandi, perché in input viene letta la matrice delle distanze, che non deve essere ricalcolata a ogni passo
- Il metodo più noto è il metodo **K-means**

Il metodo k-means

- Algoritmo che assegna le unità al gruppo con il centroide più vicino (McQueen, 1967)

1. Partizione delle unità in K gruppi iniziali
2. riallocazione di ogni unità al gruppo con il centroide più vicino e ricalcola il centroide del gruppo che ha 'perso' l'unità riallocata e del gruppo che l'ha acquisita
3. Ripete il passo 2 fino a che non ci sono più riallocazioni

- Versione alternativa: iniiziare calcolando K centroidi (semi) e quindi procedere con il passo 2.
- La partizione finale dipende molto da quella iniziale (o dai semi) in quanto la maggior parte degli spostamenti di unità avviene nei primi passi

unità	variabili	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

gruppo	Coordinate centroidi	
	m(x_1)	m(x_2)
(AB)	2	2
(CD)	-1	-2

$$\bar{x}_{jk,new} = \begin{cases} \dfrac{n_k \bar{x}_{jk} + x_{ij}}{n_k + 1} & i \textit{ aggiunta a } k \\ \dfrac{n_k \bar{x}_{jk} - x_{ij}}{n_k - 1} & i \textit{ rimossa da } k \end{cases}$$

Esempio K-means

creare K=2 gruppi omogenei

1. Dividere in maniera arbitraria le unità in 2 gruppi, p.e. (A,B) e (C,D) e calcolarne i centroidi
2. Calcolare la distanza euclidea di ciascuna unità *i* dal centroide dei due gruppi (*k*=1,2) e riassegnarla al gruppo più vicino. Ricalcolare le coordinate dei centroidi dei due gruppi dopo lo spostamento (*j*=1,..*p* variabili).

Calcoli in k_means.xls

Algoritmo k-means alternativo

- Utilizzare il criterio $\min E = \sum d^2_{i,c(k_i)}$

  che considera la somma delle distanze dal centroide del gruppo di appartenenza di tutte le possibili partizioni delle *i* unità in *K* gruppi
- Quante sono le possibili partizioni di *n* unità in *K* gruppi (num_partizioni.sas)? $\frac{1}{K!}\sum_{j=1}^{K}(-1)^{K-j}\binom{K}{j}j^n$
- Nell'esempio fatto sono 7 (4 unità in 2 gruppi)

Calcoli in k_means.xls

PROC FASTCLUS

- Costruisce una partizione delle unità in K gruppi utilizzando il Metodo *nearest centroid sorting* (Anderberg, 1973)
  - algoritmo *leader* di Hartigan (1975) per trovare valori iniziali
  - *algoritmo k-means* di MacQueen (1967) che minimizza la somma delle distanze al quadrato dal centroide di gruppo

1. Si seleziona un insieme di punti iniziali (*cluster seeds*)
2. Ogni osservazione è assegnata al seme più vicino.
3. I semi sono sostituiti dalle medie dei cluster formati al passo precedente
4. Si ripetono i passi 2 e 3 fino a che non ci sono più spostamenti di osservazioni

- L'algoritmo usa la distanza Euclidea
- Se ci sono valori mancanti su una o più variabili, PROC FASTCLUS calcola una distanza *corretta* utilizzando solo i valori validi.

Determinazione dei semi iniziali (*seeds*)

- I valori iniziali (*semi*) sono scelti tra le osservazioni senza valori mancanti
- Si può specificare il numero massimo di cluster con l'opzione MAXCLUSTERS= (default=100)
- si può specificare la distanza minima tra i semi iniziali con l'opzione RADIUS= (default=0)

Determinazione dei semi iniziali (*seeds*)

- PROC FASTCLUS seleziona la prima osservazione completa (senza missing) come primo seme.
- L'osservazione successiva che ha una distanza dalla prima maggiore o uguale a RADIUS= diventa il secondo seme.
- Le osservazioni successive sono scelte come seme iniziale se sono separate da tutti i semi precedenti per una distanza maggiore o uguale al raggio, fino a che non si raggiunge il numero massimo di semi previsto (specificato in MAXCLUSTERS=).

Determinazione dei semi iniziali

- Se un'osservazione è completa, ma non può essere un *nuovo seme*, PROC FASTCLUS valuta se questa osservazione può rimpiazzare uno dei semi già selezionati tramite 2 test:
  - Un vecchio seme è sostituito se la distanza tra l'osservazione e il seme più vicino è maggiore della distanza minima tra i semi. Il seme che verrà sostituito è scelto tra i due semi più vicini, scegliendo quello che ha la distanza più piccola con gli altri semi quando l'altro seme è rimpiazzato con l'osservazione corrente
  - Se l'osservazione non supera il primo test, si fa un altro test: l'osservazione rimpiazza il seme più vicino se la distanza più piccola tra l'osservazione e i tutti gli altri semi, escluso quello più vicino, è più grande della più piccola distanza del seme più vicino con tutti gli altri semi.
  - Se l'osservazione non supera nemmeno questo secondo test PROC FASTCLUS non sostituisce alcun seme e passa a considerare l'osservazione successiva.

DETTAGLI PROC FASTCLUS

- Il metodo di inizializzazione della procedura FASTCLUS è sensibile agli outliers.
- gli outliers spesso costituiscono gruppi con una sola unità
- FASTCLUS è programmata per archivi grandi, $n \geq 100$.
- Per data set piccoli i risultati possono essere molto sensibili all'ordinamento delle osservazioni nel data set.

Metodo k-mean servizi pubblici

- Dati raccolti su 22 imprese che hanno erogato servizi pubblici nel 1975 negli USA
  (JW Tab 12.12, dati in T12-4.DAT)
- Le variabili hanno unità di misura molto diverse➔usiamo dati standardizzati.

Utility.sas

Metodo k-mean servizi pubblici

- PROC FASTCLUS utilizza il data set standardizzato come input e crea un archivio che contiene le variabili originali e due nuove variabili: Cluster e Distance.
  - Cluster è numero del cluster cui l'osservazione è stata assegnata
  - Distance è la distanza tra l'osservazione e il centroide del gruppo di appartenenza
- Conviene replicare la procedura per diversi valori di MAXCLUSTERS= e confrontare i risultati ottenuti.

Analisi dei gruppi su dati stati americani: metodo k-means

```
proc fastclus data=statistd(drop=freddo area)
     out=out maxc=5;
where stato ne 'alaska';
id stato;
run;
```

cluster.sas

Output di FASTCLUS: cluster

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	12	0.6183	2.3609	3	3.0508
2	9	0.5452	1.8676	5	1.2974
3	10	0.6145	1.8730	5	2.0377
4	4	0.6656	2.0542	3	2.5065
5	14	0.4830	2.0349	2	1.2974

Output di FASTCLUS: variabili

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
pop75	1.00238	0.63928	0.627159	1.682109
reddito	0.90665	0.47067	0.752957	3.047884
analfab	1.00727	0.56713	0.709411	2.441282
vita	0.99589	0.58932	0.679016	2.115419
crimini	0.99842	0.51429	0.756781	3.111528
diploma	0.98012	0.62636	0.625630	1.671155
OVER-ALL	0.98240	0.57096	0.690370	2.229666

Output di FASTCLUS: medie di gruppo

```
The FASTCLUS Procedure
Replace=FULL  Radius=0  Maxclusters=5 Maxiter=1
```

Cluster Means

Cluster	pop75	reddito	analfab	vita	crimini	diploma
1	-0.05892	-1.22661	1.416384	-1.077875	1.109835	-1.356692
2	-0.20320	0.82452	-0.370046	1.078719	-0.881956	0.634559
3	0.07359	0.50075	-0.295308	-0.428786	0.409585	0.213197
4	2.53922	0.74446	-0.114842	-0.127458	0.547738	0.175436
5	-0.53482	-0.26750	-0.771081	0.656588	-0.909260	0.432338

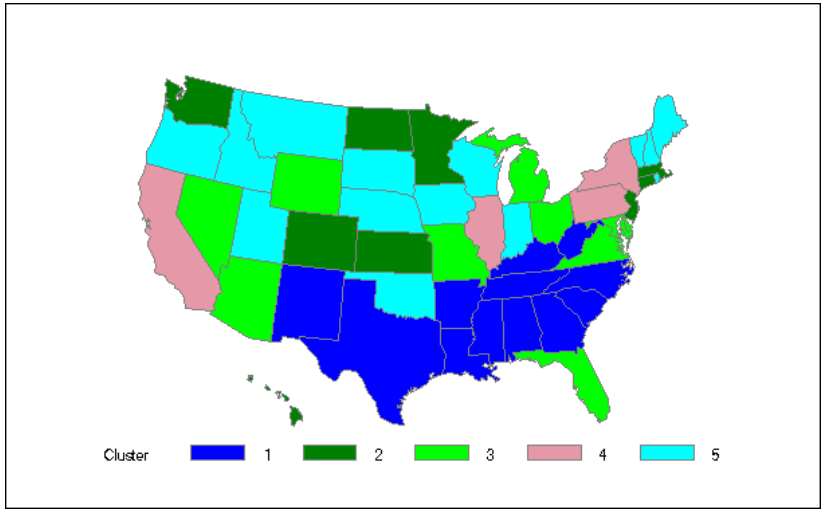
Ulteriore output di PROC FASTCLUS

- Pseudo *F* Statistic

  $$[([(R^2)/(c - 1)])/([(1 - R^2)/(n - c)])]$$

  dove R^2 è calcolato sul totale delle osservazioni, c è il numero di gruppi che formano al partizione, e n è il numero di osservazioni. Per approfondimenti vedere l'esempio 23.2 del capitolo 23, "The CLUSTER Procedure"
- R^2 (Observed Overall R-Squared), se si specifica l'opzione SUMMARY
- Approximate Expected Overall R-Squared: valore atteso (approssimato) di R^2 sotto l'ipotesi che le variabili siano incorrelate. Il valore è mancante se il numero di gruppi è maggiore di 1/5 delle osservazioni (p.e. se n=100 e c=25).
- CCC (Cubic Clustering Criterion) calcolato sotto l'ipotesi che le variabili siano incorrelate. Il valore è mancante se il numero di gruppi è maggiore di 1/5 delle osservazioni

Analisi dei gruppi con metodo k-means



Classificazione di 20 alimenti in base alla composizione

Composizione di 20 alimenti (% su 100 gr), da Zani, vol I

gruppo	alimento	acqua	proteine	lipidi	glucidi	energia
1	pane	31.0	8.1	0.5	64.0	276
1	grissini	8.5	12.3	13.9	69.0	433
1	crackers	6.0	9.4	10.0	80.1	428
1	fette bis	4.0	11.3	6.0	83.0	410
1	biscotti	2.2	6.6	7.9	85.4	418
1	pasta	12.4	10.8	0.3	82.8	356
1	riso	12.9	7.0	0.6	87.6	362
1	pizza	40.5	4.0	4.0	51.9	247
2	carote	91.6	1.1	0.0	7.6	33
2	lattuga	94.3	1.8	0.4	2.2	19
2	patate	78.5	2.1	1.0	18.0	85
2	pomodori	94.0	1.0	0.2	3.5	19
2	spinaci	90.1	3.4	0.7	3.0	31
2	zucchine	93.6	1.3	0.1	1.4	11
2	limoni	89.5	0.6	0.0	2.3	11
3	arance	87.2	0.7	0.2	7.8	34
3	banane	76.8	1.2	0.3	15.5	66
3	mele	85.6	0.2	0.3	11.0	45
3	pesche	90.7	0.8	0.1	6.1	27
3	uva	80.3	0.5	0.1	15.6	61

zani.sas

Matrice di correlazione

```
proc corr data=alimenti;
var acqua glucidi lipidi proteine energia ;
run;
```

Pearson Correlation Coefficients, N = 20

	acqua	glucidi	lipidi	proteine	energia
acqua	1.00000	-0.99215	-0.70584	-0.92340	0.99619
glucidi	-0.99215	1.00000	0.62313	0.90044	0.98271
lipidi	-0.70584	0.62313	1.00000	0.70497	0.75423
proteine	-0.92340	0.90044	0.70497	1.00000	0.93284
energia	-0.99619	0.98271	0.75423	0.93284	1.00000

Correlazioni molto elevate tra tutte le variabili!

PROC CLUSTER: legame singolo

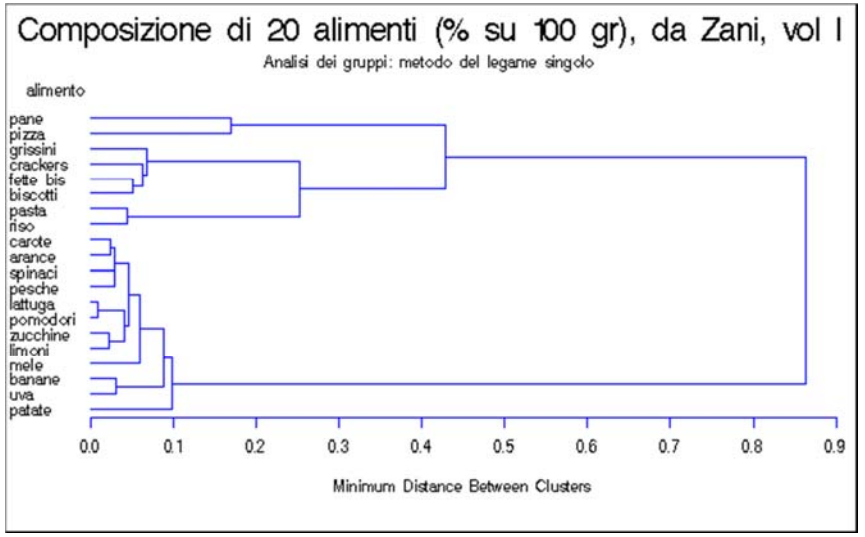
```
proc cluster data=alimenti outtree=tree
          method=single;
var acqua glucidi lipidi proteine energia;
id alimento;
run;

proc tree data=tree out=out n=3 horizontal space=2;
id alimento;
copy alimento acqua glucidi lipidi proteine energia ;
run;
```

Analisi dei gruppi: legame singolo

Cluster History

NCL	-----Clusters Joined------		FREQ	Norm Min Dist	Tie
19	lattuga	pomodori	2	0.008	
18	zucchine	limoni	2	0.0216	
17	carote	arance	2	0.0231	
16	spinaci	pesche	2	0.0292	
15	CL17	CL16	4	0.0293	
14	banane	uva	2	0.0313	
13	CL19	CL18	4	0.0411	
12	pasta	riso	2	0.0437	
11	CL15	CL13	8	0.046	
10	fette bis	biscotti	2	0.0505	
9	CL11	mele	9	0.0589	
8	crackers	CL10	3	0.0632	
7	grissini	CL8	4	0.0678	
6	CL9	CL14	11	0.0888	
5	CL6	patate	12	0.098	
4	pane	pizza	2	0.1691	
3	CL7	CL12	6	0.2517	
2	CL4	CL3	8	0.4286	
1	CL2	CL5	20	0.8635	

Dendrogramma: legame singolo



Descrizione dei gruppi

```
proc means data=out mean std cv  maxdec=2 fw=6;
class cluster;
var acqua glucidi lipidi proteine energia ;
run;
```

CLUSTER	N Obs	Variable	Mean	Std Dev	Coeff of Variation
1	12	acqua	87.68	6.14	7.00
		glucidi	7.83	5.89	75.17
		lipidi	0.28	0.30	105.17
		proteine	1.23	0.87	71.01
		energia	36.83	23.25	63.12
2	6	acqua	7.67	4.40	57.33
		glucidi	81.32	6.55	8.05
		lipidi	6.45	5.34	82.75
		proteine	9.57	2.34	24.48
		energia	401.17	33.67	8.39
3	2	acqua	35.75	6.72	18.79
		glucidi	57.95	8.56	14.76
		lipidi	2.25	2.47	109.99
		proteine	6.05	2.90	47.92
		energia	261.50	20.51	7.84

Metodo k-means, dati alimenti

```
proc fastclus data=alimenti out=out maxc=3;
      var acqua glucidi lipidi proteine energia;
       id alimento;
      run;
```

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	2	10.5201	16.6337	2	144.5
2	6	15.6840	45.8691	1	144.5
3	12	11.0787	50.0899	1	236.0

Output di k-means

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
acqua	37.91621	5.71816	0.979650	48.140786
glucidi	35.13126	6.27319	0.971471	34.052165
lipidi	3.98472	2.96605	0.504257	1.017175
proteine	4.16753	1.61137	0.866239	6.476025
energia	172.01114	26.60716	0.978592	45.711159
OVER-ALL	80.36532	12.58082	0.978073	44.606209

Cluster Means

Cluster	acqua	glucidi	lipidi	proteine	energia
1	35.7500000	57.9500008	2.2500000	6.0500002	261.5000000
2	7.6666665	81.3166669	6.4500000	9.5666667	401.1666667
3	87.6833331	7.8333334	0.2833333	1.2250000	36.8333333

Cluster Standard Deviations

Cluster	acqua	glucidi	lipidi	proteine	energia
1	6.71751442	8.55599097	2.47487373	2.89913807	20.50609665
2	4.39530034	6.54779870	5.33769602	2.34150961	33.67145181
3	6.13585994	5.88856112	0.29797295	0.86982235	23.24898173

Esempio dati HATCO (Hair et al., 1998, pp. 120-134)

- Dati provenienti da studio di segmentazione clienti della HATCO
- 100 osservazioni
- 14 variabili relative alla soddisfazione e alle caratteristiche dei clienti (aziende)

hatco.sas crea SAS Data Set

cluster3.sas crea SAS Data Set