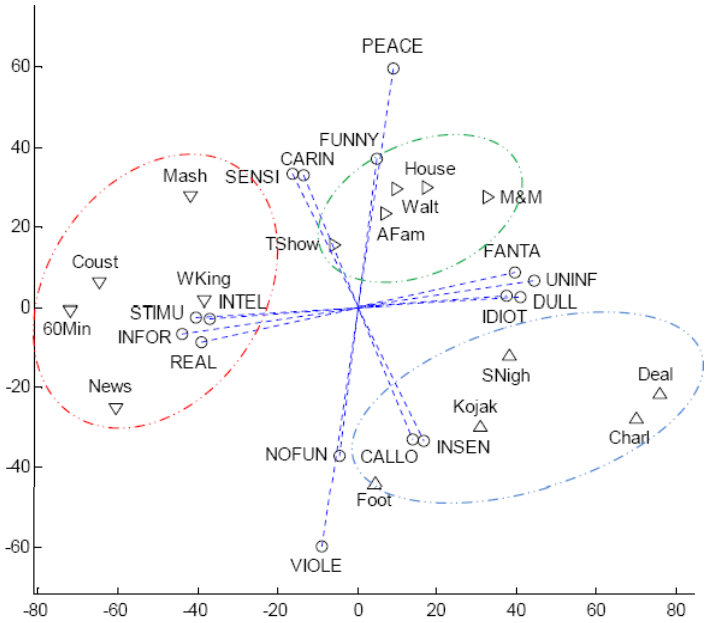


BIG DATA
Parte I - Borra

Statistica Multivariata II

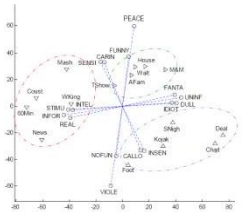
Simone Borra

Analisi dei Gruppi (AdG)

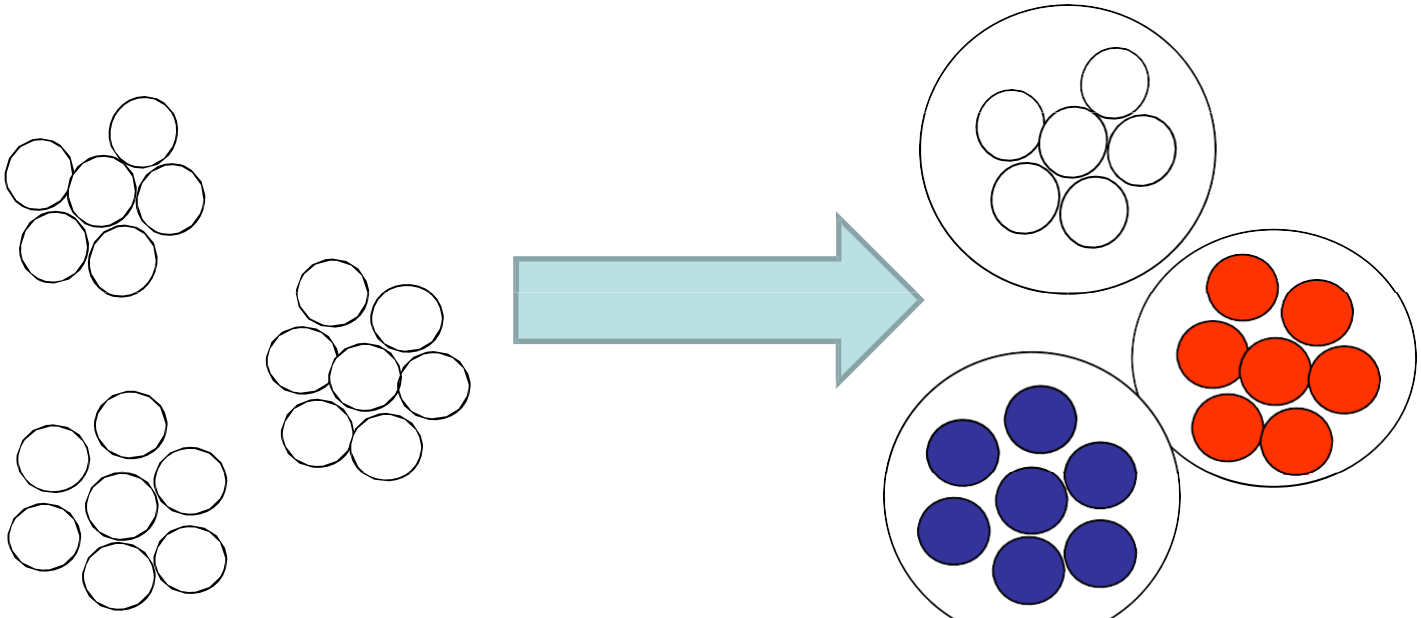
Obiettivo: raggruppare oggetti simili.

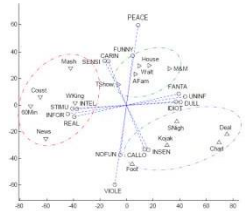
Perché classificare?

- descrivere un collettivo di individui;
- scelta delle città rappresentative per la rilevazione dei prezzi;
- individuazione di strati ai fini del campionamento;
- individuare delle tipologie di clienti per introdurre nuovi prodotti;
- segmentazione del mercato;
- indicazioni per sviluppare una nuova teoria.



Logica cluster analysis





adg: dati

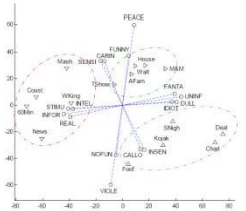
1) matrice unità×variabili

- dicotomiche
- qualitative politomiche
- quantitative

2) matrice di prossimità tra unità

- similarità
- dissimilarità

- Ci occuperemo solo del caso 1.3



adg: decomposizione della devianza

La devianza totale di *k* variabili quantitative (i.e. la somma delle devianze di ciascuna variabile) rilevate su *n* unità suddivise in *G* gruppi, può sempre decomporsi come

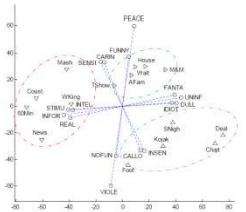
$$D_T = \sum_{l=1}^{n} d(\mathbf{x}_l, \bar{\mathbf{x}})^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} d(\mathbf{x}_{ig}, \bar{\mathbf{x}})^2$$

$$= \sum_{g=1}^{G} n_g d(\bar{\mathbf{x}}_g, \bar{\mathbf{x}})^2 + \sum_{g=1}^{G} \sum_{i=1}^{n_g} d(\mathbf{x}_{ig}, \bar{\mathbf{x}}_g)^2 = D_B + D_W$$

dove d(**a**,**b**) è la distanza euclidea tra **a** e **b**.

Nel caso di una sola variabile abbiamo

$$\sum_{l=1}^{n} (\bar{x} - x_l)^2 = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{x} - x_{ig})^2 = \sum_{g=1}^{G} n_g (\bar{x} - \bar{x}_g)^2 + \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{x}_g - x_{ig})^2$$



adg: decomposizione della devianza

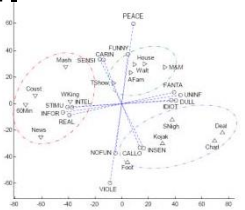
Dalle precedenti formule discende che data una partizione in gruppi delle unità del dataset la variabilità totale (misurata come **devianza totale**) si decompone nella **somma di**

Devianza Between
variabilità delle medie (indice di diversità tra gruppi)

Devianza Within
variabilità entro i gruppi (indice di disomogeneità interna dei gruppi)

Una partizione ottimale sarà quella che ha alta D_B e bassa D_W.

Importante: D_B implica bassa D_W e vice versa essendo il valore di D_T indipendente dalla partizione.



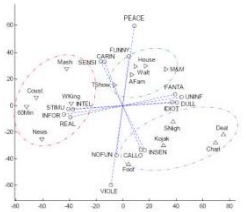
adg: K-medie

Il numero di cluster *G* è deciso a priori. Si sceglie la partizione in *G* gruppi "migliore" secondo il criterio:

$$\min D_W = \min \sum_{g=1}^{G} \sum_{i=1}^{n_g} d(\mathbf{x}_{ig}, \bar{\mathbf{x}}_g)^2$$

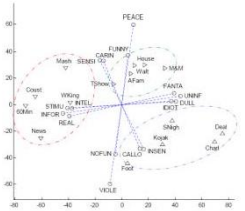
Algoritmo

1) scelgo *G* centri iniziali;
2) assegno ogni unità ad un centro secondo il criterio della minima distanza;
3) ricalcolo i *G* centri come baricentri;
4) itero 2 e 3 finchè i nuovi baricentri risultano non "significativamente" diversi dai precedenti.



adg: proprietà K-medie

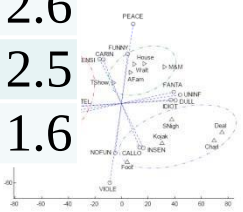
- ad ogni passo la devianza within diminuisce;
- ogni cluster è convesso, ovvero una unità del cluster *g* non può essere una media ponderata delle unità appartenenti al cluster $k \neq g$.
- L'algoritmo spesso raggiunge solo degli ottimi locali quindi il risultato finale dipende fortemente dalla scelta dei *G* centri iniziali. Tale fenomeno si acuisce quando non esistono dei cluster ben separati. Generalmente si cerca di ovviare a tale inconveniente ripetendo l'algoritmo con differenti centri di partenza oppure utilizzando particolari procedure di scelta dei centri iniziali.



adg: esempio K-medie

Componenti del latte espresse in percentuale

	Acqua	Proteine	Grassi	Lattosio
Cavallo	90.1	2.6	1.0	6.9
Asino	90.3	1.7	1.4	6.2
Mulo	90.0	2.0	1.8	5.5
Cammello	87.7	3.5	3.4	4.8
Lama	86.5	3.9	3.2	5.6
Zebra	86.2	3.0	4.8	5.3
Pecora	82.0	5.6	6.4	4.7
Bufalo	82.1	5.9	7.9	4.7
Maiale G.	81.9	7.4	7.2	2.7
Volpe	81.6	6.6	5.9	4.9
Maiale	82.8	7.1	5.1	3.7
Coniglio	71.3	12.3	13.1	1.9
Topo	72.5	9.2	12.6	3.3
Daino	65.9	10.4	19.7	2.6
Renna	64.8	10.7	20.3	2.5
Balena	64.8	11.1	21.2	1.6



adg: esempio K-medie

Componenti del latte espresse in percentuale

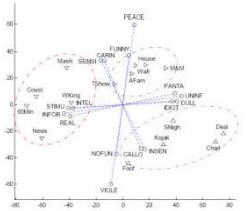
Variable	Mean	Std Dev	Minimum	Maximum
acqua	80.0	9.2	64.8	90.3
proteine	6.4	3.5	1.7	12.3
grassi	8.4	6.9	1	21.2
lattosio	4.2	1.6	1.6	6.9

A causa delle diverse intensità operiamo la standardizzazione delle variabili

Applichiamo una K-medie con K=5
Ottenendo I seguenti valori medi e dev.stand. per singolo gruppo

Cluster Means				
Cluster	stnd_acqua	stnd_proteine	stnd_grassi	stnd_lattosio
1	1.00	-1.11	-0.96	1.17
2	0.42	-0.43	-0.40	0.44
3	-1.62	1.23	1.74	-1.22
4	-0.89	1.23	0.64	-0.99
5	0.25	0.23	-0.33	-0.61

Cluster Standard Deviations				
Cluster	stnd_acqua	stnd_proteine	stnd_grassi	stnd_lattosio
1	0.20	0.28	0.14	0.40
2	0.31	0.45	0.25	0.16
3	0.07	0.10	0.11	0.34
4	0.09	0.63	0.05	0.62
5	0.07	0.06	0.22	0.44



adg: esempio K-medie

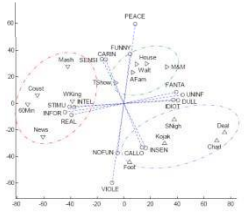
Componenti del latte espresse in percentuale

Cluster Means				
Cluster	**stnd_acqua**	**stnd_proteine**	**stnd_grassi**	**stnd_lattosio**
1	1.00	-1.11	-0.96	1.17
2	0.42	-0.43	-0.40	0.44
3	-1.62	1.23	1.74	-1.22
4	-0.89	1.23	0.64	-0.99
5	0.25	0.23	-0.33	-0.61

Poichè le variabili sono standardizzate la media delle variabili per l'intero campione è 0

Unità	**Gruppo**
cavallo	1
asino	1
mulo	1
lama	1
cammello	2
zebra	2
pecora	2
bufalo	2
volpe	2
daino	3
renna	3
balena	3
coniglio	4
topo	4
maiale_g	5
maiale	5

Cluster Means				
Cluster	**stnd_acqua**	**stnd_proteine**	**stnd_grassi**	**stnd_lattosio**
1	+	-	-	+
2	=	=	=	+
3	-	+	+	-
4	-	+	+	-
5	+	+	=	-



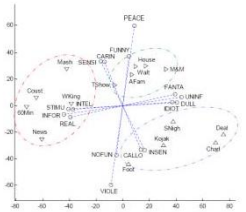
adg: metodi gerarchici

Un metodo gerarchico, fornisce *n*-1 (*n* = numero di unità) partizioni in *n*, *n*-1, …, *G*, …,2 gruppi.

La partizione in *G* gruppi è ottenuta da quella in *G*+1 fondendo i due gruppi più “simili”.

Si basano sul principio che se due unità in una partizione in *G* gruppi appartengono ad uno stesso cluster allora devono appartenere ad uno stesso cluster anche in una partizione in *H* gruppi, per *H* = *G*-1,…,2.

I metodi gerarchici differiscono tra loro solo per il criterio secondo cui fondono due gruppi in ogni step.



adg: metodi gerarchici

Ward La fusione avviene minimizzando l'incremento di devianza within.

Metodi del legame La fusione avviene tra i due gruppi più vicini (i.e. con distanza minore). Nei tre metodi del legame che consideriamo la distanza tra cluster è definita come:

Singolo $$d(C_h, C_k) = \min_{i \in C_h, i \in C_k} d_{ij}$$

Medio $$d(C_h, C_k) = \frac{1}{n_h n_k} \sum_{i \in C_h, i \in C_k} d_{ij}$$

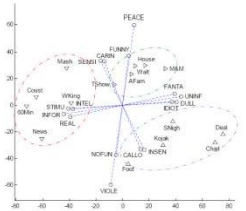
Completo $$d(C_h, C_k) = \max_{i \in C_h, i \in C_k} d_{ij}$$

dove d_{ij} è la distanza tra la i-ma e la j-ma unità

adg: esempio legame singolo

	a	b	c	d	e
a	0				
b	1	0			
c	48	12	0		
d	6	20	63	0	
e	35	48	20	2	0

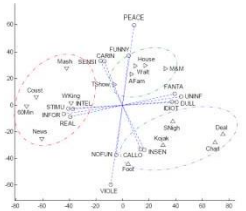
	a,b	c	d	e
a,b	0			
c	12	0		
d	6	63	0	
e	35	20	2	0



adg: esempio legame singolo

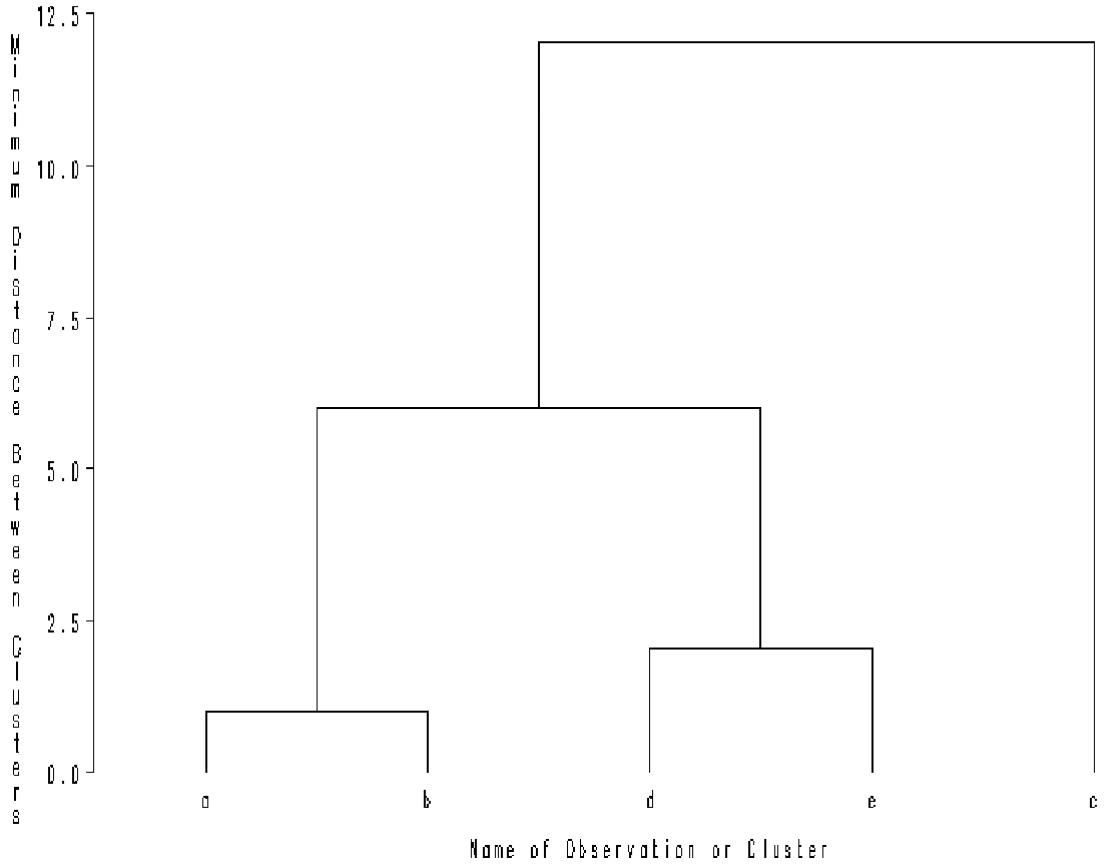
	a,b	c	d,e
a,b	0		
c	12	0	
d,e	6	20	0

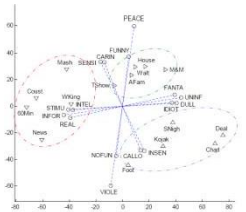
	a.b.d.e	c
a,b,d,e	0	
c	12	0



adg: esempio legame singolo

Dendrogramma





adg: confronto tra metodi gerarchici

Ward

Particolarmente efficace quando i gruppi hanno la stessa numerosità e matrice di varianze e covarianze.

Legame singolo

Invariante per trasformazioni monotone crescenti della matrice delle distanze iniziale;

Cerca di minimizzare la lunghezza degli “anelli”;

Utile per individuare gruppi irregolari (generalmente non convessi);

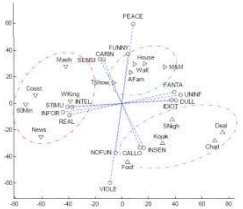
Legame medio

Tende a produrre gruppi aventi la stessa varianza.

Legame completo

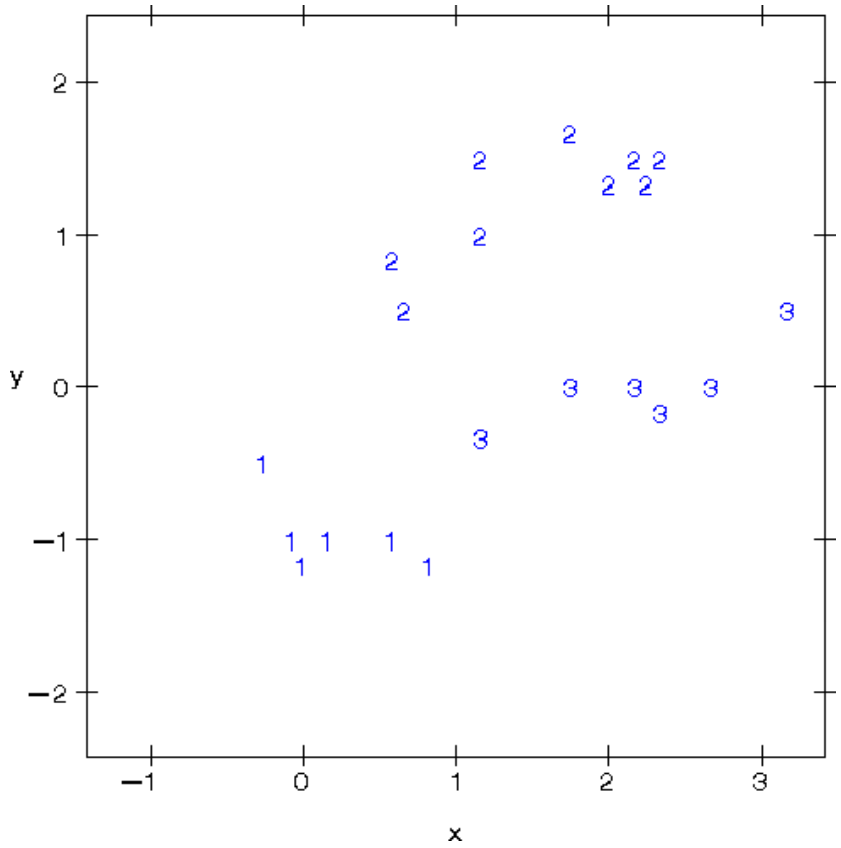
Invariante per trasformazioni monotone crescenti della matrice delle distanze iniziale;

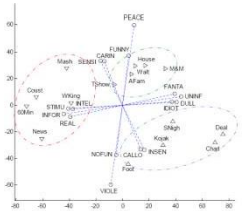
Tende a produrre gruppi aventi lo stesso diametro.



adg: confronto tra metodi gerarchici

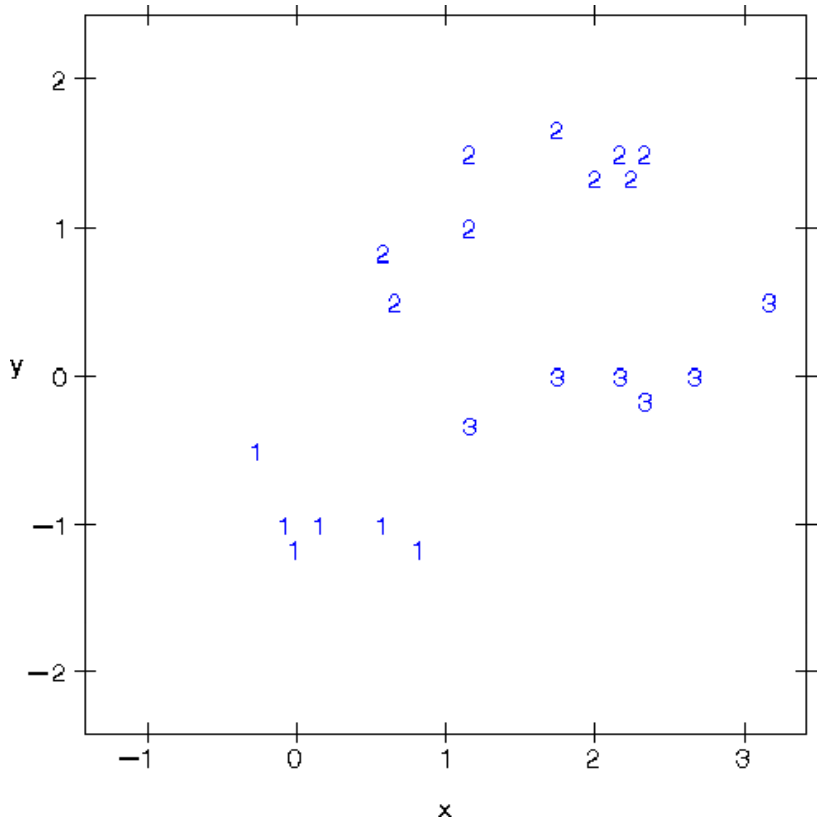
Ward (2,3)

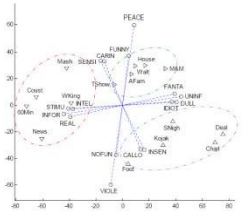




adg: confronto tra metodi gerarchici

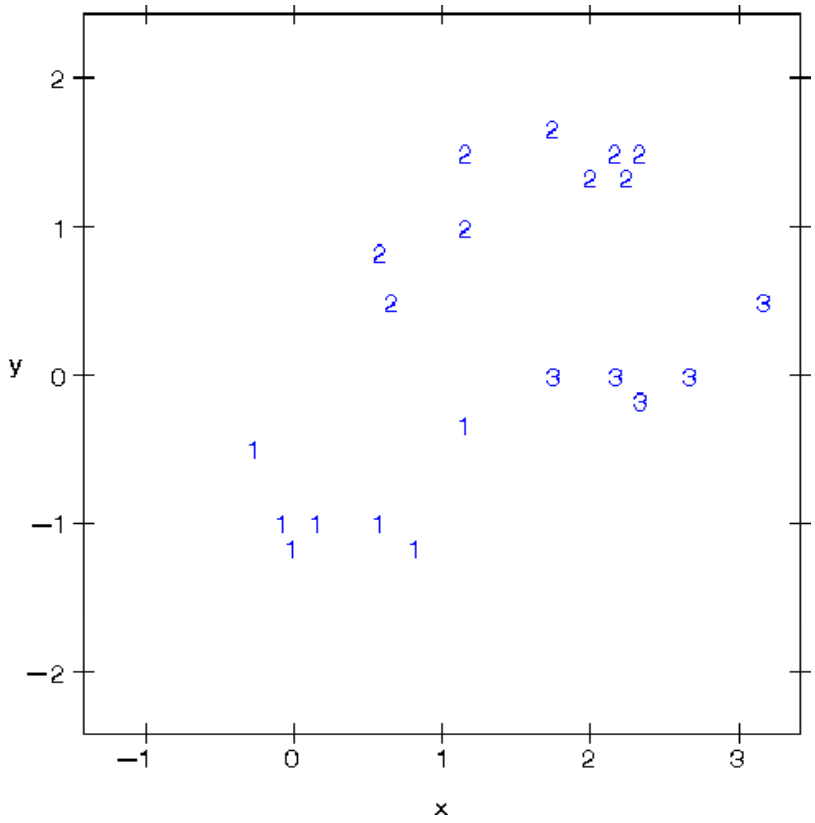
Singolo (1,3)

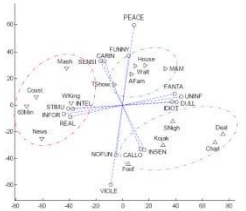




adg: confronto tra metodi gerarchici

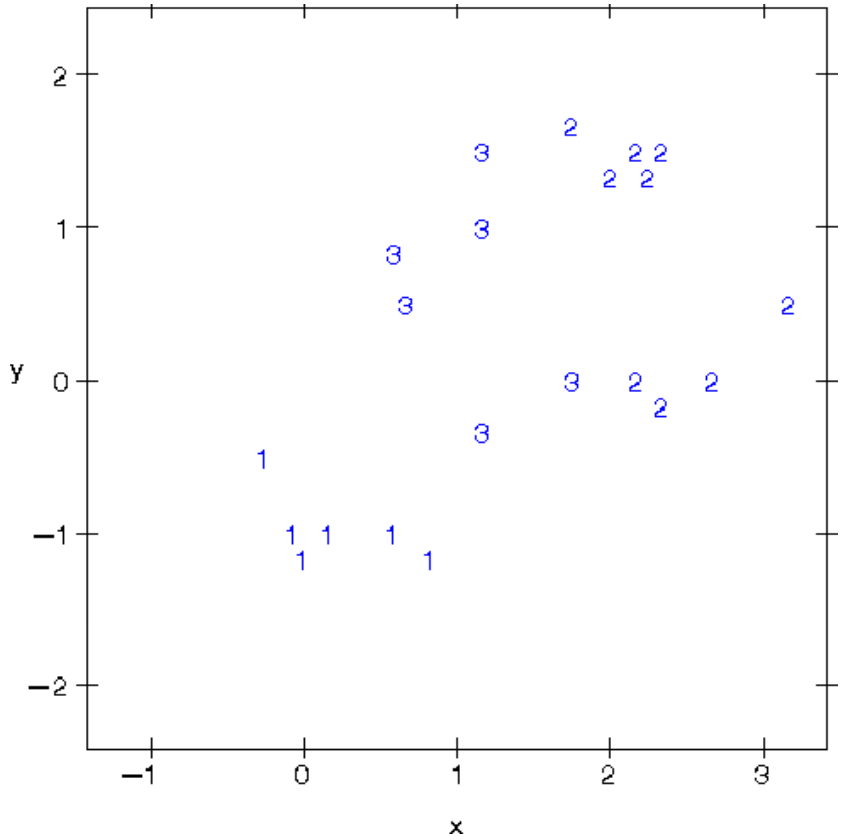
Medio (2,3)

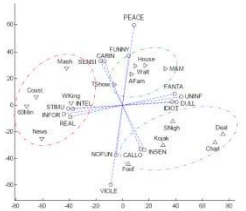




adg: confronto tra metodi gerarchici

completo (2,3)





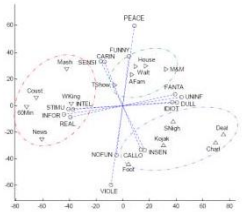
adg: scelta del numero di gruppi

Pseudo *F*: $$pF = \frac{D_B / (G-1)}{D_W / (n-G)}$$

È un indice di qualità della partizione. Se *pF* decresce in modo monotono al diminuire di *G* allora non c'è evidenza di una particolare struttura. Altrimenti si sceglie la partizione corrispondente al massimo (locale).
(Si suppongono cluster sferici).

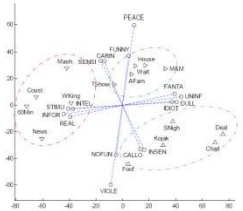
Pseudo T^2: $$pT^2 = \frac{D_{C_m} - (D_{C_h} + D_{C_k})}{(D_{C_h} + D_{C_k}) / (n_h + n_k - 2)}$$

Si utilizza solo per i metodi gerarchici. misura la perdita di informazione dovuta alla fusione dei gruppi C_h e C_k, di numerosità, rispettivamente, n_h e n_k, nel gruppo C_m.
E' la *pF* calcolata sulla bi-partizione ottenuta considerando solo le unità che appartengono ai due cluster che sono stati fusi. Ovviamente il suo valore dovrebbe essere piccolo. Viene utilizzata generalmente in congiunzione con la *pF*.



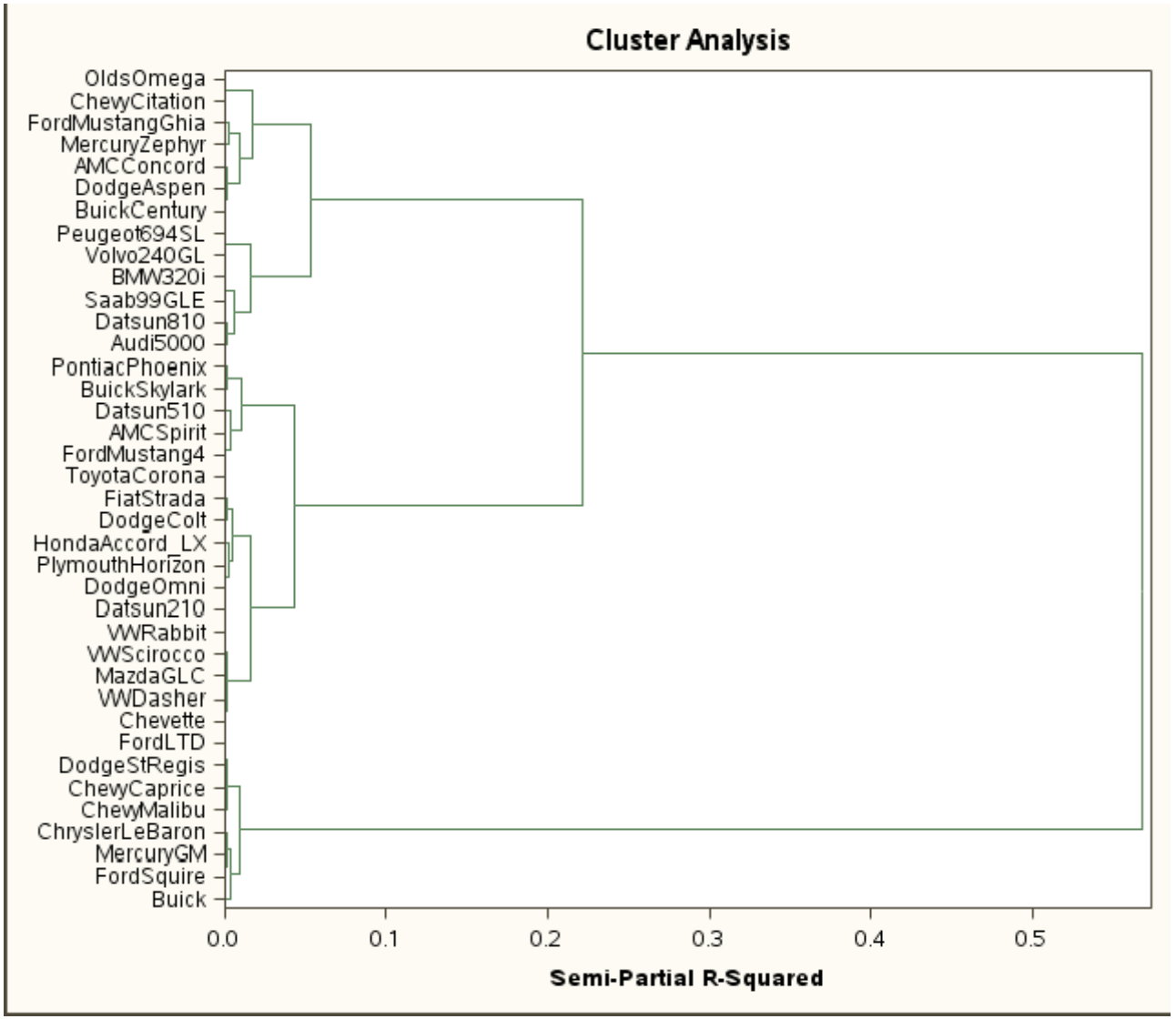
adg: scelta del numero di gruppi

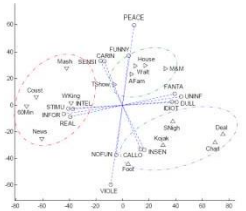
- Utilizzo gli indici *pF* e/o pT^2
- Taglio del dendrogramma
- Informazioni a priori
- Interpretabilità
- Forme forti
  Individuare gruppi di unità che vengono classificate nello stesso gruppo da più metodi



adg: esempio Ward

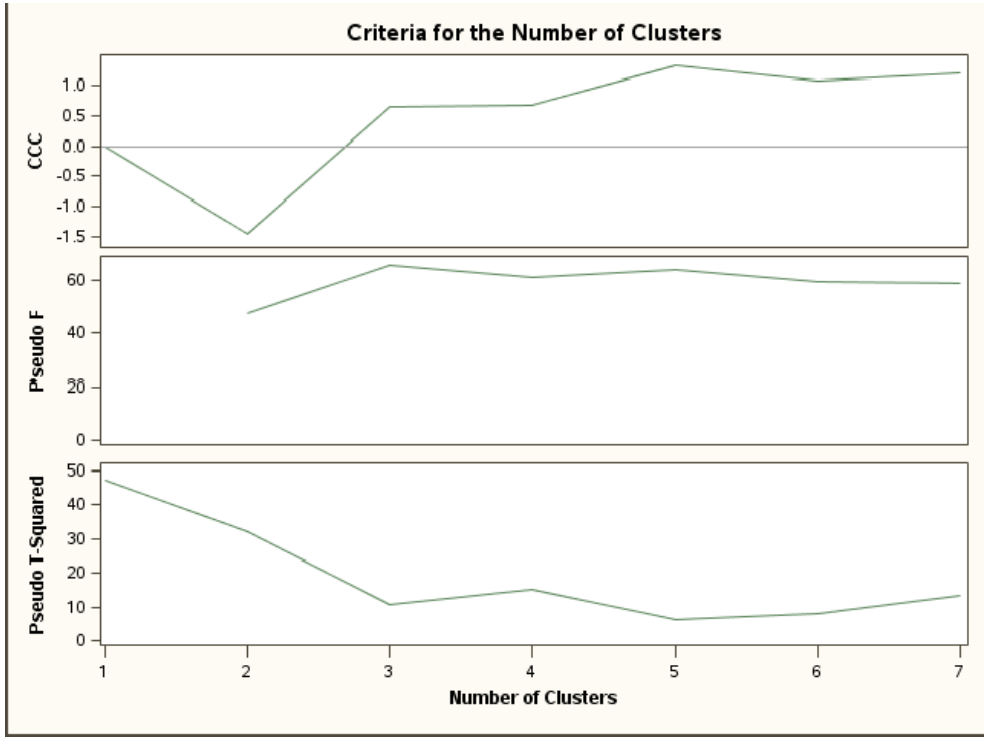
Si considerano le variabili standardizzate: consumo, peso, drive, cavalli, grandezza, cilindri

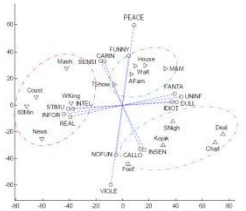




Pseudo-T^2 e pseudo-F

N. Clusters			Freq	Pseudo F	Pseudo t-squared
15	MercuryZe	FordMusta	2	81.2	.
14	Buick	CL21	4	76.5	5
13	CL24	Datsun510	4	73.8	8.1
12	CL16	CL23	5	72.2	3.5
11	CL19	CL31	4	70	5.6
10	CL14	CL17	8	64.1	7.1
9	CL22	CL15	5	61.6	6.8
8	CL13	CL18	6	61.1	6.1
7	CL20	CL12	11	58.5	13.5
6	CL11	CL26	6	59.3	8.2
5	CL9	CL30	7	**63.9**	**6.1**
4	CL	CL	17	60.	1
3	CL6	CL5	13	**65.6**	**10.7**
2	CL4	CL3	30	47.4	32
1	CL10	CL2	38	.	47.4





adg: esempio Ward

ANOVA per verificare la capacità discriminante di ogni variabile

consumo

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	2	29.6492	14.8246	70.59	<.0001

peso

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	2	31.21646	15.60823	94.46	<.0001

drive

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	2	17.19853	8.599267	15.2	<.0001

Cavalli

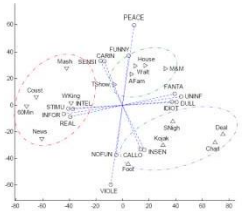
Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	2	30.71859	15.3593	85.58	<.0001

grandezza

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	2	32.22248	16.11124	118.03	<.0001

cilindri

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	2	34.24601	17.12301	217.61	<.0001

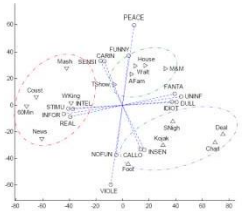


adg: esempio Ward

Confronto tra valori medi

CLUSTER	Mean of zconsumo	Mean of zpeso	Mean of zdrive	Mean of zcavalli	Mean of zgrandezza	Mean of zcilindri
1	0.96	-0.87	0.41	-0.91	-0.77	-0.87
2	-1.12	1.49	-1.30	1.37	1.67	1.63
3	-0.56	0.22	0.26	0.35	-0.02	0.14
	Std. Dev.	Std. Dev.	Std. Dev.	Std. Dev.	Std. Dev.	Std. Dev.
1	0.47	0.37	0.77	0.40	0.25	0.00
2	0.18	0.32	0.32	0.40	0.37	0.00
3	0.54	0.49	0.90	0.46	0.48	0.48

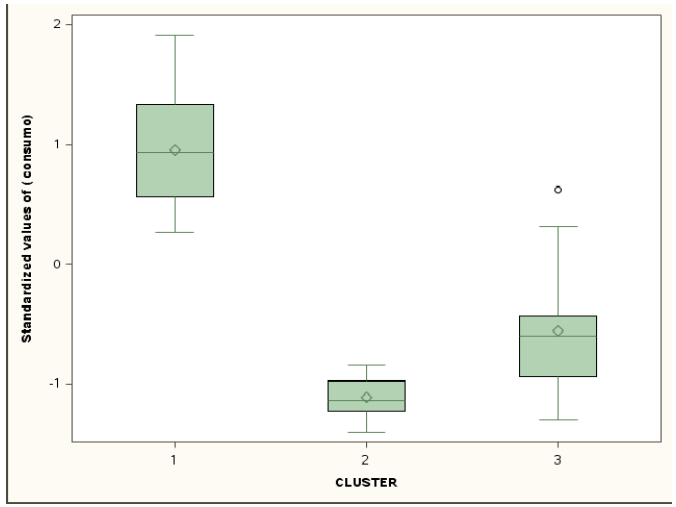
Primo gruppo: macchine utilitarie, economiche, city car, di piccole dimensioni
Secondo gruppo: macchine di grossa cilindrata, berline o sportive, di lusso
Terzo gruppo: macchine di media dimensione



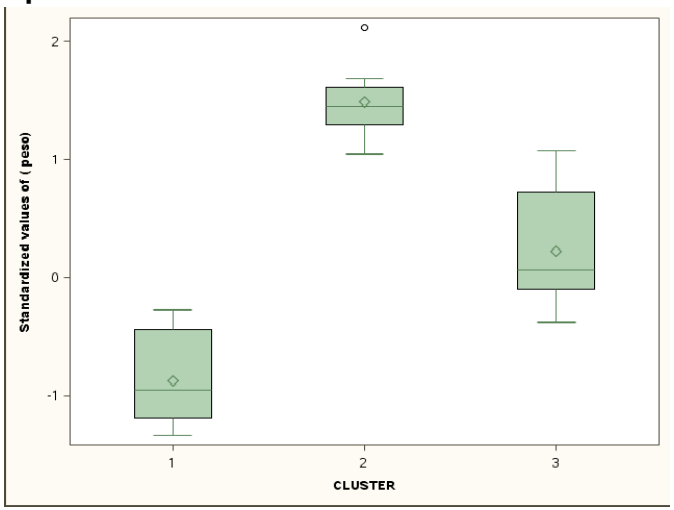
adg: esempio Ward

Box-plot per i tre diversi gruppi

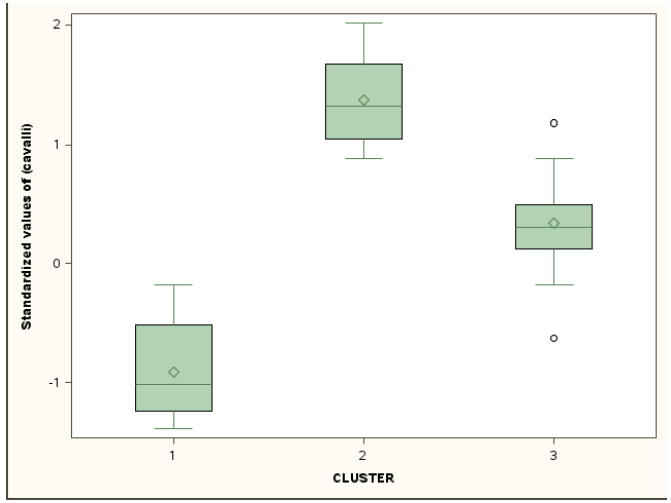
consumo

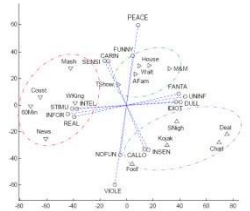


peso



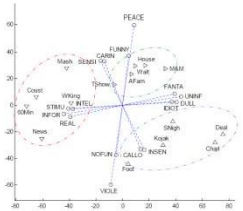
cavalli





adg: estensioni

I metodi di cluster considerati utilizzano la distanza euclidea come misura della diversità, o dissimilarità, tra unità statistiche.
Possiamo generalizzare l'uso dei metodi introdotti riferendoci in generale a misure di dissimilarità diverse dalla distanza euclidea.
In generale, possiamo basare un metodo di cluster su una qualunque misura di prossimità tra unità o oggetti, la quale può essere:
-direttamente rilevata (es.: "confusion matrices");
-calcolata in base a delle variabili (qualitative e/o quantitative).



Analisi in Componenti Principali

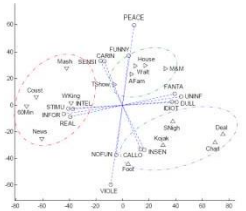
Input: *J* variabili quantitative rilevate su *n* unità.

Output

- Rappresentazione grafica:
  delle unità; delle variabili; unità e variabili.
- Variabili sintesi.

Obiettivo Investigare le relazioni:

- di dissimilarità tra unità
- di associazione (correlazione) tra le variabili
- unità-variabili

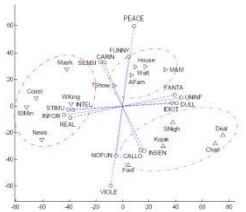


ACP: esempio

12 unità (distretti censuari della città di Los Angeles),

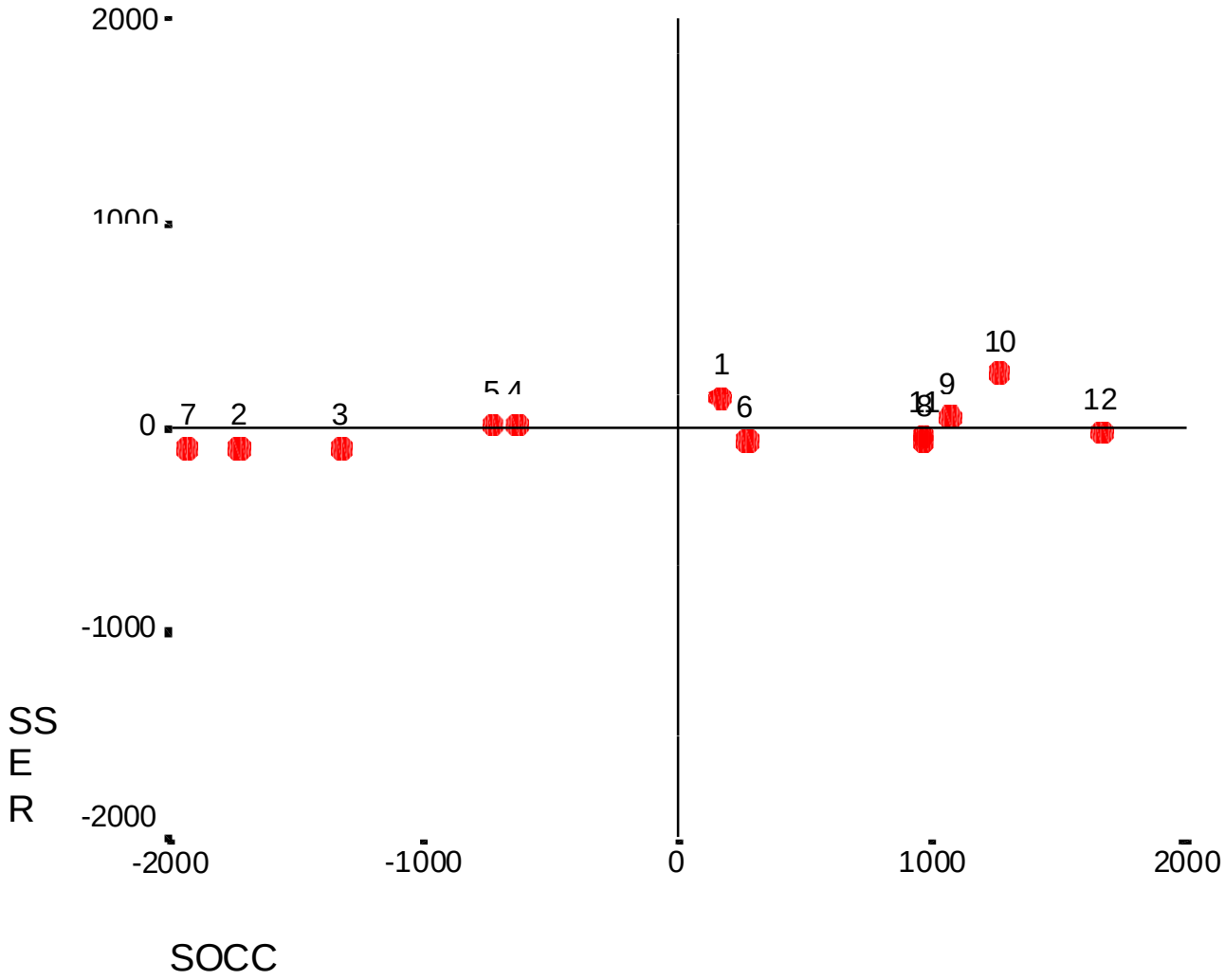
5 variabili socio-economiche:

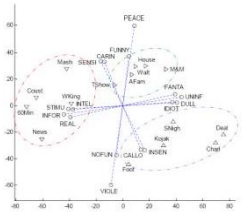
	Pop.	Scuo.	Occ.	Serv.	VCase
1	5700	12.8	2500	270	25000
2	1000	10.9	600	10	10000
3	3400	8.8	1000	10	9000
4	3800	13.6	1700	140	25000
5	4000	12.8	1600	140	25000
6	8200	8.3	2600	60	12000
7	1200	11.4	400	10	16000
8	9100	11.5	3300	60	14000
9	9900	12.5	3400	180	18000
10	9600	13.7	3600	390	25000
11	9600	9.6	3300	80	12000
12	9400	11.4	4000	100	13000



ACP: esempio

Variabili scarto





ACP: esempio

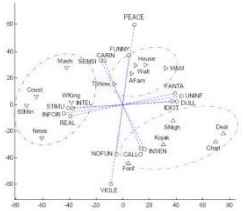
La distanza

$$d(\underset{\mathbf{x}}{p}, \underset{\mathbf{x}}{q}) = \sqrt{\left(x_{p,occ} - x_{q,occ}\right)^2 + \left(x_{p,ser} - x_{q,ser}\right)^2}$$

costituisce un indice di dissimilarità tra unità il quale però è maggiormente influenzato dalle variabili che hanno un'alta varianza. Infatti

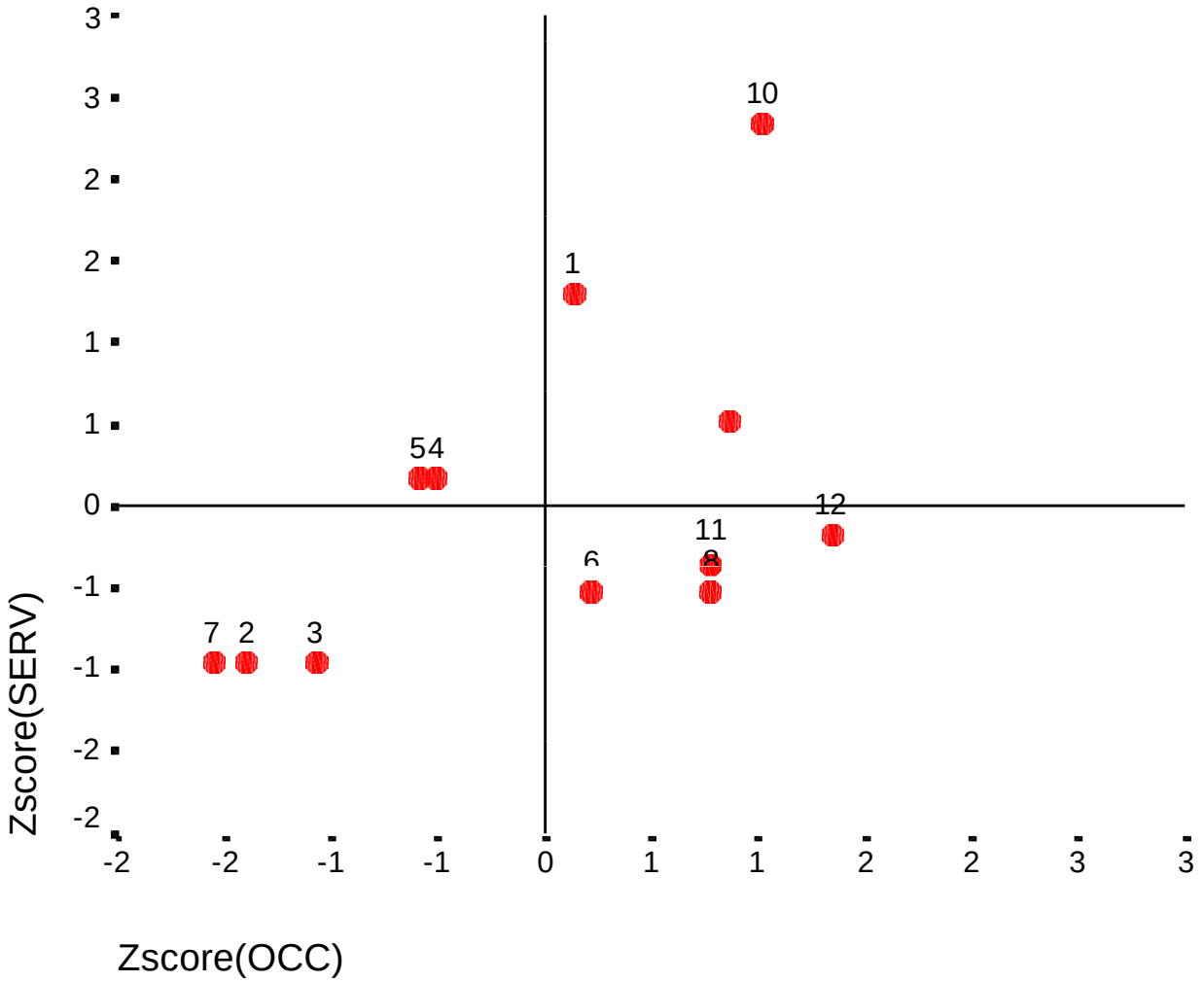
$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$$

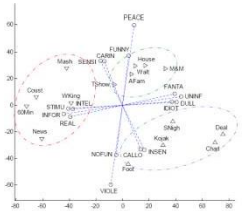
Per eliminare questo effetto standardizziamo le variabili, i.e. dividiamo ogni variabile scarto (X - μ) per la sua deviazione standard.



ACP: esempio

Variabili standardizzate





ACP: definizione

Problema Individuare delle nuove variabili, a due a due incorrelate e combinazione lineare delle variabili originarie, in grado di rappresentare “al meglio” le distanze tra le unità.

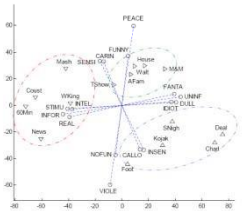
Soluzione Calcolo delle nuove variabili $\mathbf{c}_1, \ldots, \mathbf{c}_J$ (dette componenti principali) tali che

$$Var(\mathbf{c}_k) = \max$$

sotto i vincoli $Cor(\mathbf{c}_k, \mathbf{c}_h) = 0, \quad h = 1,\ldots,k-1$

$$\mathbf{c}_k = a_{k1}\mathbf{z}_1 + \ldots + a_{kJ}\mathbf{z}_J$$

$$\sum_{j=1}^{J} a_{ki}^2 = 1$$



ACP: proprietà

- hanno media zero
- varianza decrescente: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J$
- la varianza totale (i.e. la somma delle varianze delle variabili) è uguale alla somma delle varianze delle componenti
- le distanze tra unità calcolate con le nuove variabili sono uguali alle distanze calcolate con le variabili osservate
- si dimostra che per individuare le componenti dobbiamo risolvere l'equazione

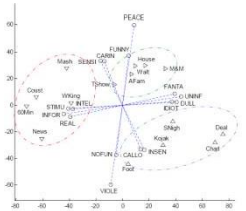
  $$\mathbf{R}\mathbf{a}_k = \lambda_k \mathbf{a}_k$$

  dove **R** è la matrice delle correlazioni tra le variabili.

ACP: esempio

Correlation Matrix					
	pop	**sch**	**occ**	**serv**	**house**
pop	1	0.0098	0.9724	0.4389	0.0224
sch	0.0098	1	0.1543	0.6914	0.8631
occ	0.9724	0.1543	1	0.5147	0.1219
serv	0.4389	0.6914	0.5147	1	0.7777
house	0.0224	0.8631	0.1219	0.7777	1

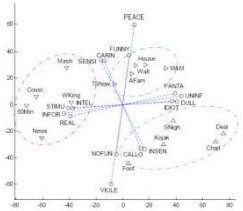
Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	2.87331359	1.0766535	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.934
3	0.21483689	0.11490283	0.043	0.977
4	0.09993405	0.08467868	0.02	0.9969
5	0.01525537		0.0031	1



ACP: selezione componenti

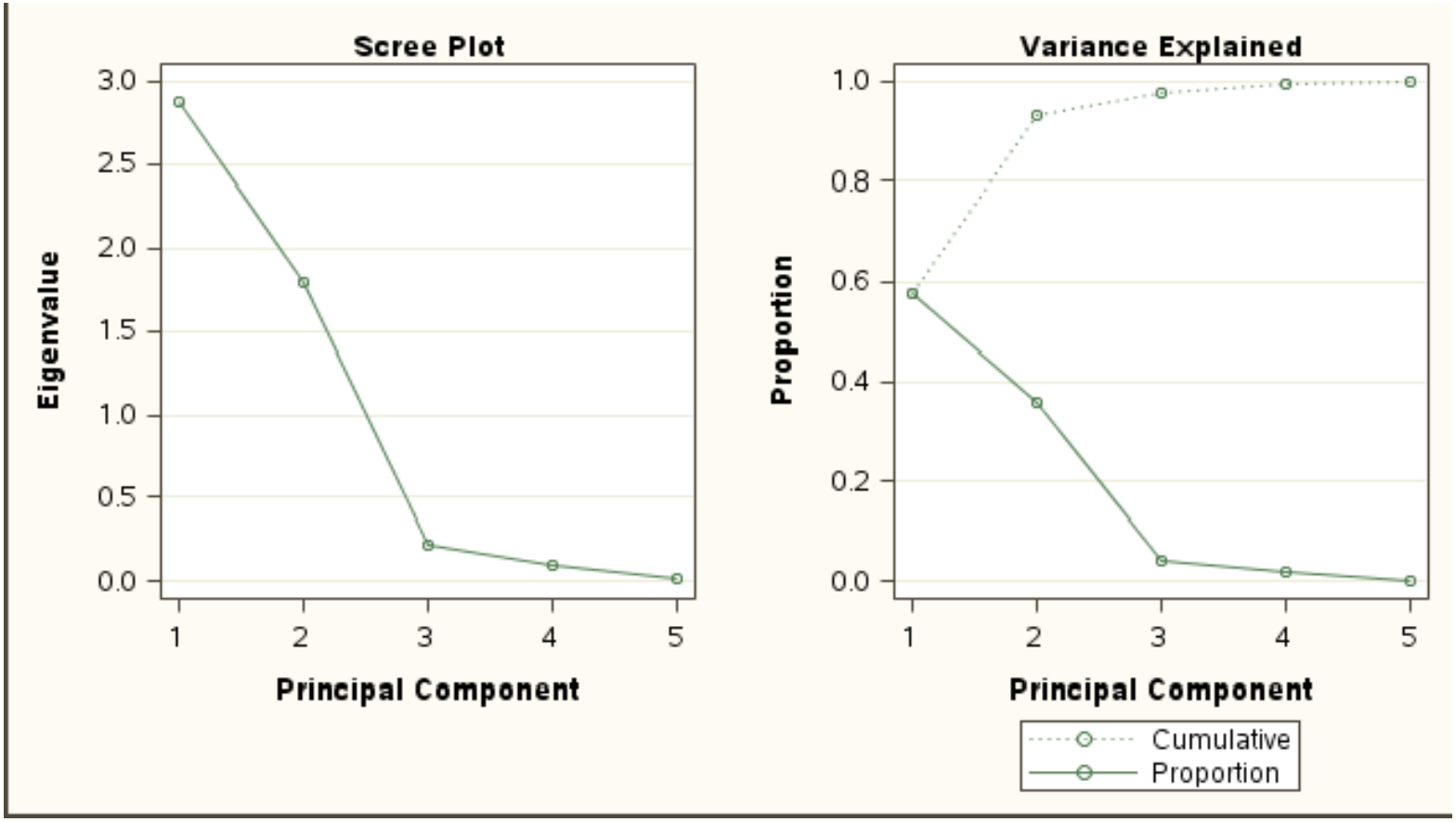
Se trascuro le componenti che hanno varianza piccola le distanze rimangono quasi inalterate. Il numero di componenti è scelto

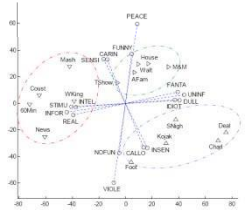
- **sulla base della percentuale di varianza spiegata**
  Estraggo un numero di componenti in grado di spiegare una quota “soddisfacente” della variabilità totale
- **trascurando le componenti con varianza inferiore a uno**
  Trascuro le componenti che hanno un contenuto informativo inferiore a quello di una singola variabile originale
- **considerando l’interpretabilità**
  Estraggo solo le componenti a cui riesco ad attribuire un significato
- **in base allo scree plot**
  Estraggo componenti fino a quando non vi è un salto significativo nella varianza spiegata



ACP: esempio

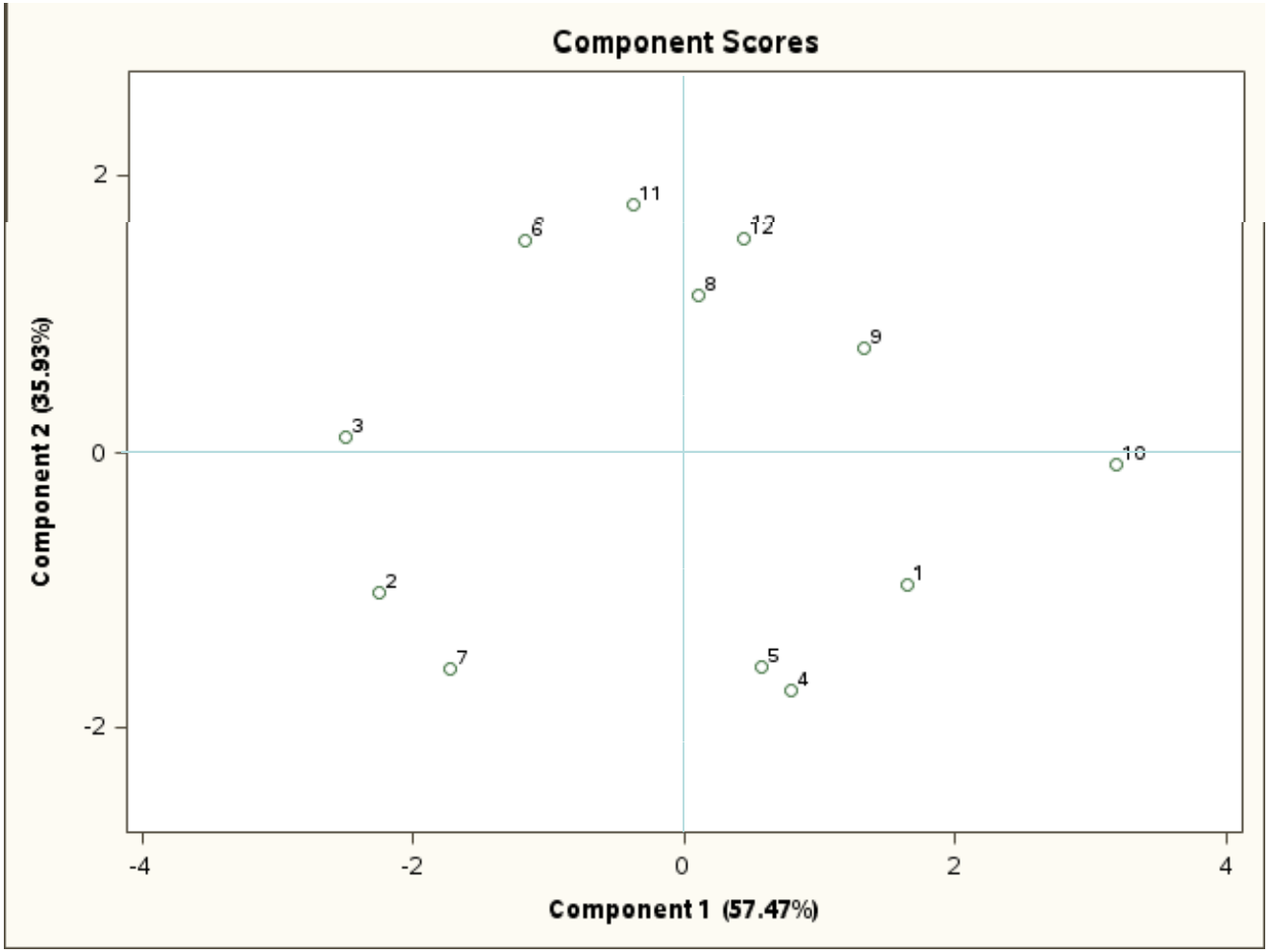
Scree Plot

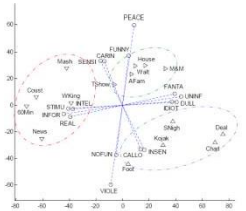




ACP: esempio

Primo piano principale





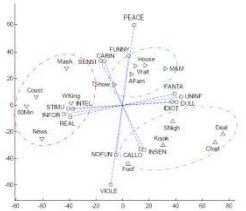
ACP: interpretazione

Problema come interpreto le distanze in relazione alle variabili originarie?
Soluzione calcolo le correlazioni tra le variabili originarie e le componenti.

Si dimostra che è possibile scrivere le variabili originali come combinazione lineare delle componenti (tutte).
Scriviamo quindi le variabili originali come

$$\mathbf{z}_j = b_{j1}\mathbf{c}_1 + b_{j2}\mathbf{c}_2 + \ldots + b_{jJ}\mathbf{c}_J$$

dove $\mathbf{c}_k = \mathbf{c}_k / \sqrt{\lambda_k} = \mathbf{c}_k / \sqrt{Var(\mathbf{c}_k)}$ sono le componenti standardizzate.



ACP: interpretazione

- $\hat{\mathbf{z}} = b_{j1}\mathbf{c}_1 + b_{j2}\mathbf{c}_2 + \ldots + b_{jK}\mathbf{c}_K$ è la regressione lineare

(minimi quadrati ordinari) della *i*-ma variabile sulle prime *K* componenti

- $b_{jk} = Cor(\mathbf{z}_j, \mathbf{c}_k)$

- $b_{jk} = \sqrt{\lambda_k} a_{jk} \Rightarrow b_{1k}^2 + b_{2k}^2 + \ldots + b_{Jk}^2 = \lambda_k$

- $Var(\hat{\mathbf{z}}_j) = Var(b_{j1}\mathbf{c}_1 + b_{j2}\mathbf{c}_2 + \ldots + b_{jK}\mathbf{c}_K)$
  $= b_{j1}^2 + \ldots + b_{jK}^2$ (Comunalità)

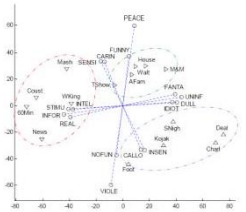
- $Cov(\hat{\mathbf{z}}_j, \hat{\mathbf{z}}_h) = b_{j1}b_{h1} + b_{j2}b_{h2} + \ldots + b_{jK}b_{hK}$

ACP: esempio

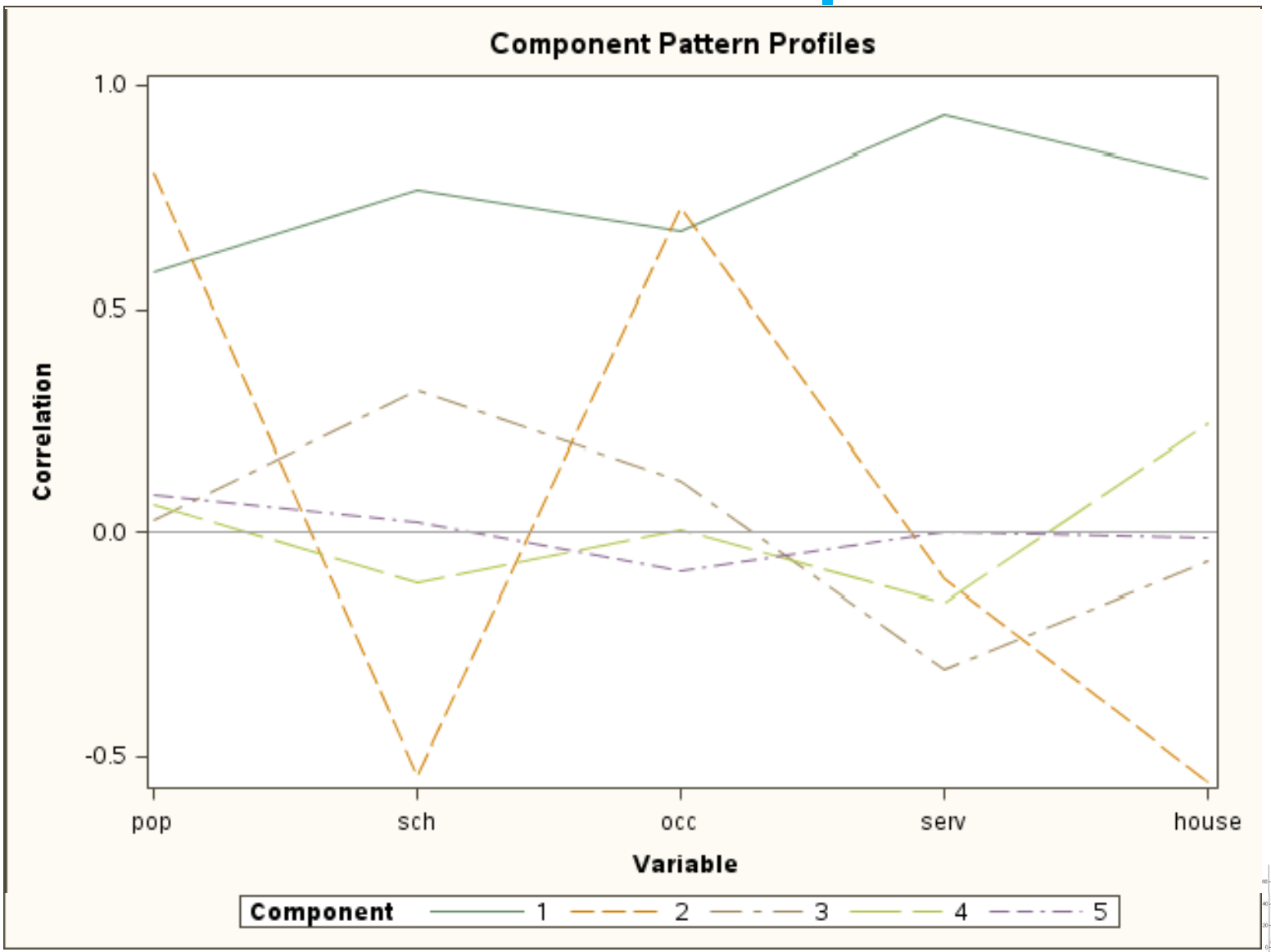
Factor Pattern		
	Factor1	Factor2
pop	0.58096	0.80642
sch	0.76704	-0.54476
occ	0.67243	0.72605
serv	0.93239	-0.10431
house	0.79116	-0.55818

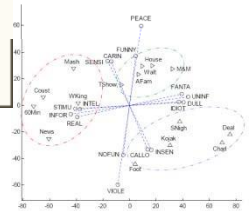
Explained by Factor	
Factor1	Factor2
2.87331	1.79666

Final Communality Estimates: Total = 4.669974				
pop	sch	occ	serv	house
0.98782629	0.88510555	0.97930583	0.88023562	0.93750041



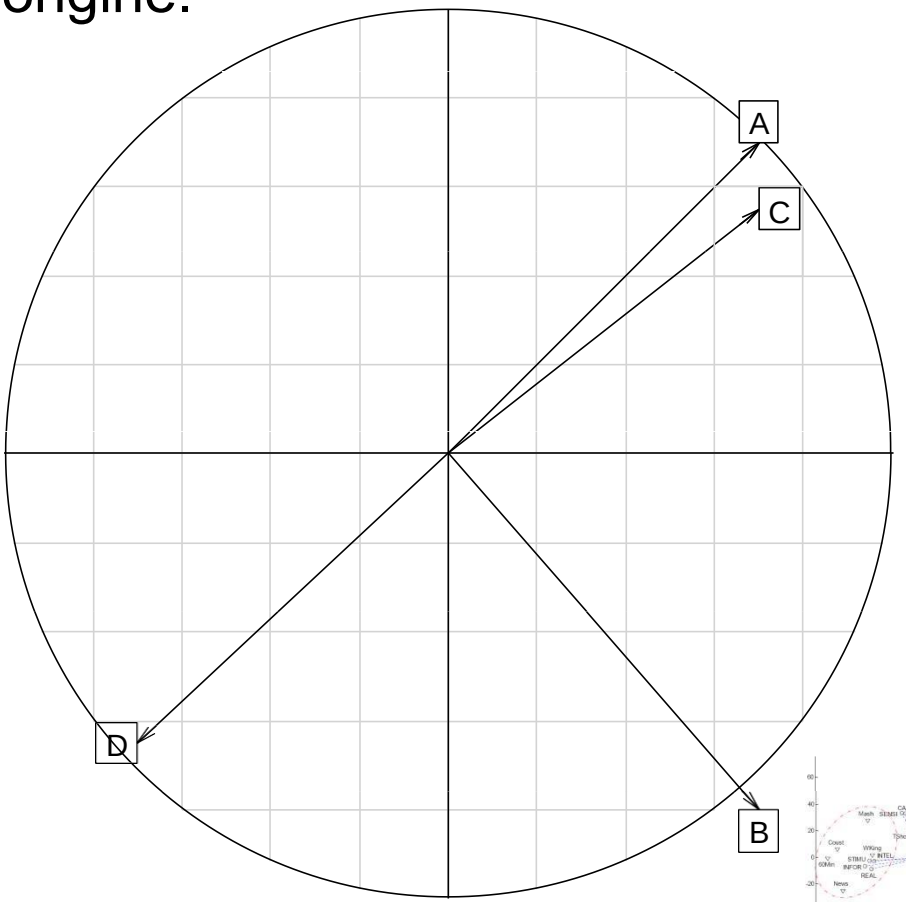
ACP: esempio


Component Pattern Profiles
1.0
0.5
0.0
-0.5
Correlation
pop
sch
occ
serv
house
Variable
Component
1
2
3
4
5



ACP: cerchio delle correlazioni

- Rappresentiamo le variabili come punti su di un piano aventi come coordinate le correlazioni che queste hanno con le componenti.
- La comunalità di una variabile è rappresentata dalla distanza al quadrato del punto dall'origine.
- La correlazione tra due variabili è rappresentata dal prodotto tra la radice del prodotto delle comunalità ed il coseno dell'angolo compreso.

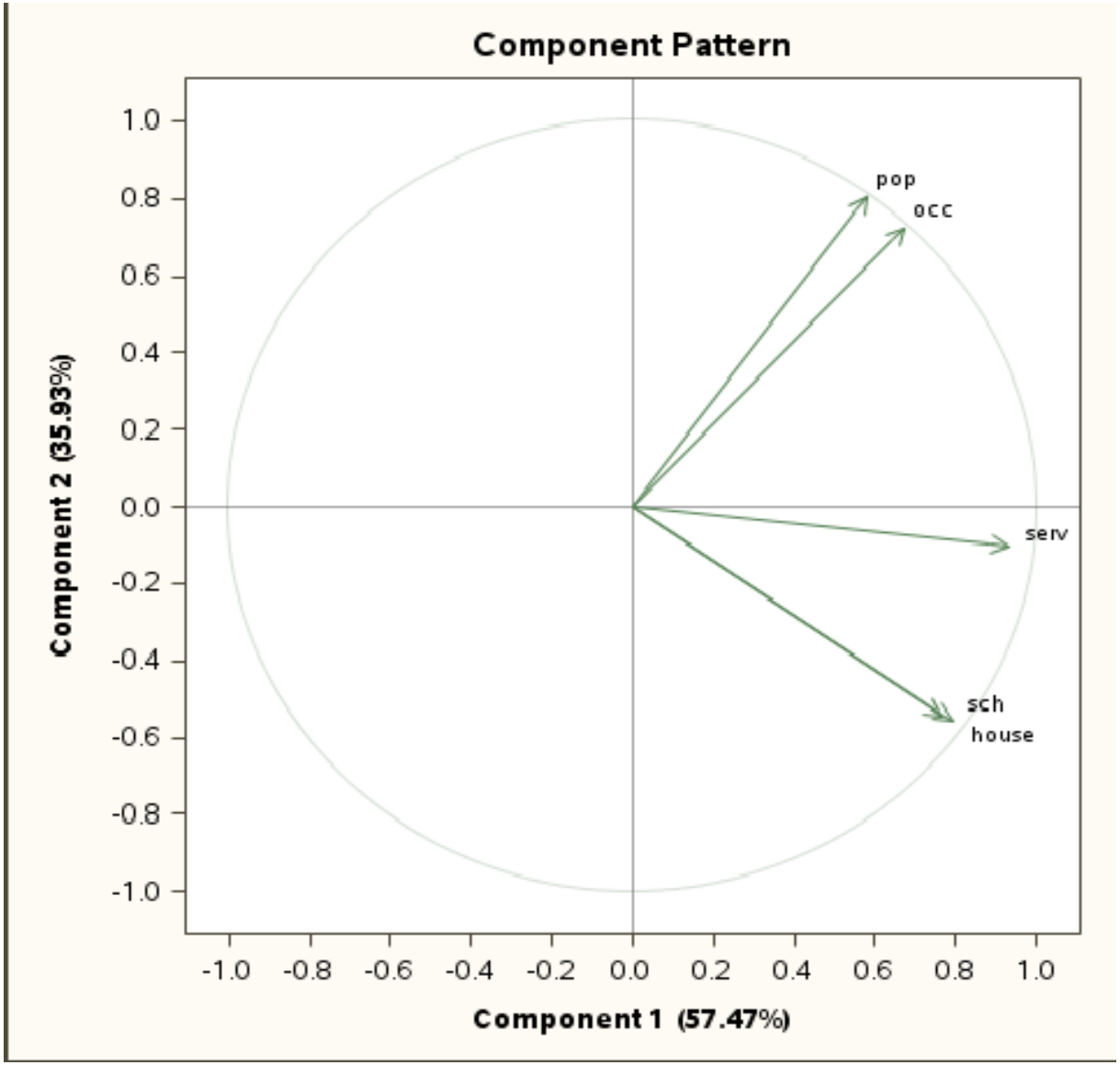


Cor(A,B) ≈ 0

Cor(A,C) ≈ 1

Cor(A,D) ≈ -1

ACP: esempio


Component Pattern
pop
occ
serv
sch
house
Component 2 (35.93%)
Component 1 (57.47%)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0