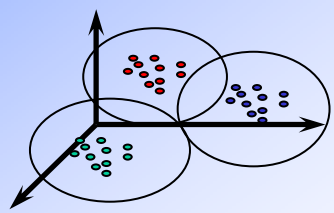
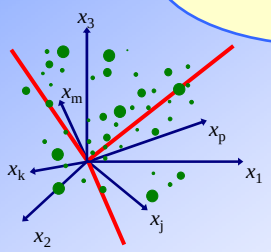
L'Analisi Multidimensionale dei Dati

Una Statistica da vedere



Analisi generale

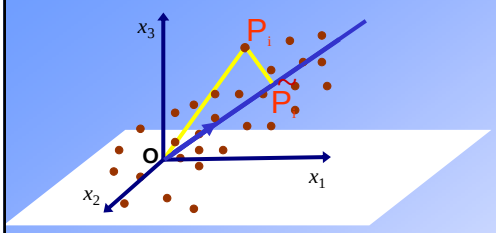
Analisi in Componenti Principali (ACP)



Analisi delle Corrispondenze (ACB e ACM)

• Analisi Generale

Metodo grafico



u=[1 0 1]

$||u||^2 = 1^2 + 0^2 + 1^2 = 2$

vettore normalizzato

$$u = \left[\frac{1}{\sqrt{2}} \quad 0 \quad \frac{1}{\sqrt{2}}\right]$$

Proiezione di un punto su **u**

P_i=[3 2 5] ➡

$$O\tilde{P}_i = [3 \quad 2 \quad 5] \times \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix} = \frac{3}{\sqrt{2}} + 0 + \frac{5}{\sqrt{2}} = 5{,}65$$

Metodo analitico

$$\sum_i \left(P_i - \tilde{P}_i\right)^2 = \min \quad \Leftrightarrow \quad \sum_i \left(O\tilde{P}_i\right)^2 = \max$$

• Analisi Generale

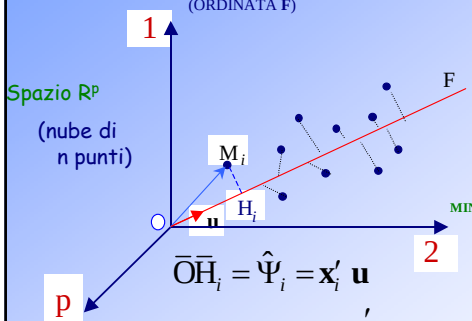
$\mathbf{X}_{(n,p)} =$

$\left\{ \begin{array}{l} \text{punti } M_i \\ \text{linee } x_i' \end{array} \right\}$

$= \hat{\Psi}_{(n,1)}$

$\hat{\Psi}_i$

Ridurre $\mathbf{X}$ ($n \times p$ *numeri*) *alla conoscenza dell'ascissa* $\hat{\Psi}_i$ (n *numeri*) *sull'asse* $\mathbf{u}$ (p *numeri*)



$$\overline{OH}_i = \hat{\Psi}_i = \mathbf{x}_i' \, \mathbf{u}$$

$$\sum_i \overline{OH}_i^2 = (\mathbf{X}\,\mathbf{u})' \, (\mathbf{X}\,\mathbf{u})$$

$$\sum_i \overline{OM}_i^2 = \text{costante}$$

$$\sum_i \overline{OM}_i^2 = \sum_i \overline{OH}_i^2 + \sum_i \overline{MH}_i^2$$

MINIMI QUADRATI : $\underset{(\mathbf{u})}{Min} \left\{ \sum_i \overline{MH}_i^2 \right\} \Leftrightarrow \underset{(\mathbf{u})}{Max} \left\{ \sum_i \overline{OH}_i^2 \right\}$

$$\boxed{\begin{array}{ll} \underset{(u)}{Max} & \{\mathbf{u}'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u}\} \\ con & : \quad \mathbf{u}'\mathbf{u} = 1 \end{array}}$$

• Analisi Generale

$$\boxed{\begin{array}{ll} \underset{(u)}{Max} & \{\mathbf{u}'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u}\} \\ con & : \quad \mathbf{u}'\mathbf{u} = 1 \end{array}}$$

LAGRANGIANA (λ moltiplicatore) L= $\mathbf{u}'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u} - \lambda\,(\mathbf{u}'\,\mathbf{u} - 1)$

CONDIZIONE DI ESTREMO

$$\left. \begin{array}{l} \dfrac{\partial}{\partial\,\mathbf{u}}\,\mathbf{u}'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u} = 2\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u} \\ \dfrac{\partial}{\partial\,\mathbf{u}}\,\lambda\,(\mathbf{u}'\,\mathbf{u} - 1) = 2\,\lambda\,\mathbf{u} \end{array} \right\} \quad \boxed{\begin{array}{l} \mathbf{X}'\,\mathbf{X}\,\mathbf{u} = \lambda\,\mathbf{u} \\ \mathbf{u}: \textit{autovettore} \\ \lambda: \textit{autovalore di } (\mathbf{X}'\,\mathbf{X}) \end{array}}$$

CONDIZIONE DI MASSIMO

$$Max\,\{\mathbf{u}'\,(\mathbf{X}'\,\mathbf{X}\;u)\} \Rightarrow Max\,\{\lambda\,\mathbf{u}'\,\mathbf{u}\} \Rightarrow Max\{\lambda\}$$

DIAGONALIZZARE $\underset{(p,p)}{(\mathbf{X}'\,\mathbf{X})}$ $\left\{ \begin{array}{l} \lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0 \\ \mathbf{u}_1 \;;\; \mathbf{u}_2 \;;\; \ldots \;;\; \mathbf{u}_p \;(\textit{normalizzati}) \end{array} \right.$

PROIETTARE gli n punti di Rp sulla retta F passante per u_1 autovettore di X'X associato al più grande autovalore λ_1

COORDINATE degli n punti su F: $\boxed{\hat{\Psi}_1 = \mathbf{X}\,\mathbf{u}_1} \Rightarrow \hat{\Psi}_1'\,\hat{\Psi}_1 = \mathbf{u}_1'\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u}_1 = \lambda_1$

• Analisi Generale

APPROSSIMAZIONE PER UN PIANO PASSANTE PER L'ORIGINE

Il piano contiene u_1

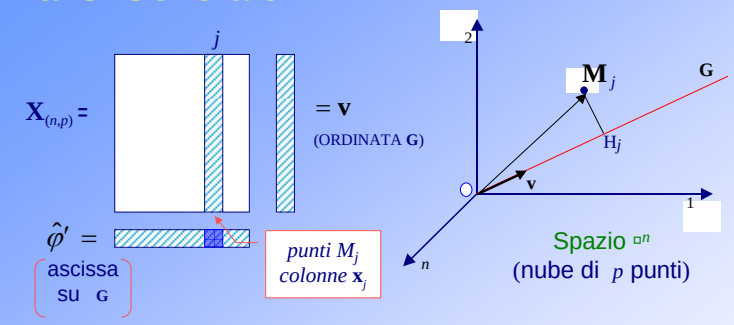
$$\begin{cases} Max\{\mathbf{u}'\ \mathbf{X}'\ \mathbf{X}\ \mathbf{u}\} \\ \mathbf{u}'\mathbf{u}=1 \\ \mathbf{u}'\mathbf{u}_1=0 \end{cases}$$

DIAGONALIZZARE $\underset{(p,p)}{(\mathbf{X}'\mathbf{X})}$ $\begin{cases} \lambda_1 > \lambda_2 > \dots > \lambda_p \geq 0 \\ \mathbf{u}_1\ ;\ \mathbf{u}_2\ ;\ \dots\ ;\ \mathbf{u}_p\ (normalizzati) \end{cases}$

SOTTO-SPAZIO DI APPROSSIMAZIONE A q DIMENSIONI

una base ortonormale del sotto-spazio vettoriale a q dimensioni che si adatta al meglio (minimi quadrati) alla nube, e' costituita da q autovettori che corrispondono ai q piu' grandi autovalori di $\mathbf{X}'\ \mathbf{X}$

• Analisi Generale



DEFINIZIONE DELL'ORDINATA G

$$\underset{(\mathbf{v})}{Min}\left\{\sum_j \overline{M}_j\overline{H}_j^2\right\} \Leftrightarrow \underset{(\mathbf{v})}{Max}\left\{\sum_j \overline{OH}_j^2\right\} \Rightarrow \left.\begin{array}{l} \overline{OH}_j=\mathbf{x}_j'\ \mathbf{v} \\ \sum_j \overline{OH}_j^2=(\mathbf{X}'\mathbf{v})'\ (\mathbf{X}'\mathbf{v}) \end{array}\right\} \boxed{\begin{array}{ll} \underset{(u)}{Max} & \{\mathbf{v}'\mathbf{X}\ \mathbf{X}'\ \mathbf{v}\} \\ con & :\ \mathbf{v}'\mathbf{v}=1 \end{array}}$$

SOTTO-SPAZIO DI APPROSSIMAZIONE A q DIMENSIONI

una base ortonormale del sotto-spazio vettoriale a q dimensioni che si adatta al meglio (minimi quadrati) alla nube, e' costituita da q autovettori $\mathbf{v}_1\ ;\ \mathbf{v}_2\ ;\ \dots\ ;\mathbf{v}_q$
che corrispondono ai q piu' grandi autovalori $\mu_1 \geq \mu_2 \geq \dots \geq \mu_q$
della matrice (n x n) simmetrica $\mathbf{X}\ \mathbf{X}'$

• Analisi Generale

RELAZIONE TRA L'APPROSSIMAZIONE IN R^p E IN R^n

Asse fattoriale $\mathbf{u}_\alpha$ in $\square^p$: $\mathbf{X}'\,\mathbf{X}\,\mathbf{u}_\alpha = \lambda_\alpha\,\mathbf{u}_\alpha$

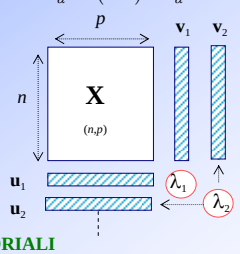
Asse fattoriale $\mathbf{v}_\alpha$ in $\square^n$: $\mathbf{X}\,\mathbf{X}'\,\mathbf{v}_\alpha = \mu_\alpha\,\mathbf{v}_\alpha$

$$\left.\begin{array}{l}\mathbf{X}'\,\mathbf{X}\,(\mathbf{X}'\,\mathbf{v}_\alpha) = \mu_\alpha\,(\mathbf{X}'\,\mathbf{v}_\alpha)\\ \mathbf{X}'\,\mathbf{X}\,\mathbf{u}_\alpha = \lambda_\alpha\,\mathbf{u}_\alpha\end{array}\right] \Rightarrow \left\{\begin{array}{l}\lambda_\alpha = \mu_\alpha\\ \mathbf{u}_\alpha = c^{nte} \times \mathbf{X}'\,\mathbf{v}_\alpha\end{array}\right.$$

$$\mathbf{u}'_\alpha\,\mathbf{u}_\alpha = (c^{nte})^2\,\mathbf{v}'_\alpha\,X\,X'\,\mathbf{v}_\alpha = (c^{nte})^2\,\lambda_\alpha = 1 \quad\Rightarrow\quad c^{nte} = \frac{1}{\lambda_\alpha}$$

RELAZIONI DI TRANSIZIONE

$$\left\{\begin{array}{l}{}_{(p,1)}\mathbf{u}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}}\,\mathbf{X}'\,\mathbf{v}_\alpha\\[2ex] {}_{(n,1)}\mathbf{v}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}}\,\mathbf{X}\,\mathbf{u}_\alpha\end{array}\right.$$



ESPRESSIONI DELLE COORDINATE SUGLI ASSI FATTORIALI

- *n* PUNTI LINEA su $\mathbf{u}\alpha$: $\hat{\psi}_\alpha = \mathbf{X}\,\mathbf{u}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha$
- *p* PUNTI COLONNA su $\mathbf{v}\alpha$: $\hat{\varphi}_\alpha = \mathbf{X}'\,\mathbf{v}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{u}_\alpha$

$$\left(\hat{\psi}'_\alpha\,\hat{\psi}_\alpha = \hat{\varphi}'_\alpha\,\hat{\varphi}_\alpha = \lambda_\alpha\right)$$

Nota per i calcoli

1) DIAGONALIZZARE $\mathbf{A} = \mathbf{X}'\,\mathbf{X} \Rightarrow (\lambda_\alpha\,;\,\mathbf{u}_\alpha)_{\alpha=1,\dots,p}$

2) CALCOLARE $\hat{\psi}_\alpha = \mathbf{X}\,\mathbf{u}_\alpha$ E $\hat{\varphi}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{u}_\alpha$

• Analisi Generale

RICOSTRUZIONE ESATTA

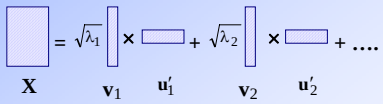
$$(\hat{\psi}_\alpha =)\ \mathbf{X}\,\mathbf{u}_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha \Rightarrow \mathbf{X}\left\{\sum_{\alpha=1}^{p}\mathbf{u}_\alpha\,\mathbf{u}'_\alpha\right\} = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha\,\mathbf{u}'_\alpha \Rightarrow \mathbf{X}\,\mathbf{u}_\alpha\,\mathbf{u}'_\alpha = \sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha\,\mathbf{u}'_\alpha$$

MA

$$\underset{(p,p)}{\mathbf{U}} = [\mathbf{u}_1|\mathbf{u}_2|\dots|\mathbf{u}_p]$$

$$\sum_{\alpha=1}^{p}\mathbf{u}_\alpha\,\mathbf{u}'_\alpha = \mathbf{U}\,\mathbf{U}' = \mathbf{I}_p$$

$$\Rightarrow \underset{(n,p)}{\mathbf{X}} = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\,\underset{(n,1)}{\mathbf{v}_\alpha}\,\underset{(1,p)}{\mathbf{u}'_\alpha}$$



Decomposizione in "valori singolari" *(Eckart-Young, 1936-1939)*

RICOSTRUZIONE APPROSSIMATA

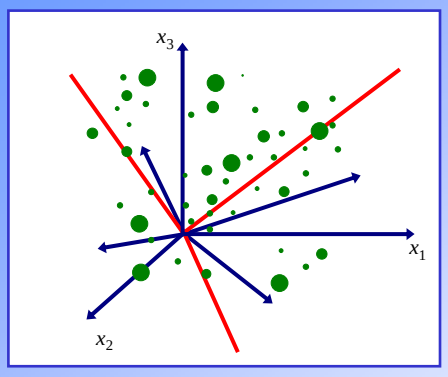
$$\underbrace{\lambda_1 > \lambda_2 > \dots > \lambda_q}_{} > \underbrace{\lambda_{q+1} > \dots > \lambda_p}_{} > 0$$

Valori giudicati "trascurabili"

$$\underset{(n,p)}{\mathbf{X}} \cong \mathbf{X}^* = \sum_{\alpha=1}^{q}\sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha\,\mathbf{u}'_\alpha$$

$$\left.\begin{array}{l}\sum_i \overline{OM}_i^2 = \sum_i\sum_j x_{ij}^2 = \text{traccia}\,(\mathbf{X}'\,\mathbf{X}) = \sum_{\alpha=1}^{p}\lambda_\alpha\\ \sum_i \overline{OH}_i^2 = (\mathbf{X}\,\mathbf{u}_1)'\,(\mathbf{X}\,\mathbf{u}_1) = \mathbf{u}'_1\,\mathbf{X}'\,\mathbf{X}\,\mathbf{u}_1 = \lambda_1\end{array}\right\}\ \frac{\lambda_1}{\sum_\alpha\lambda_\alpha} \Rightarrow \text{SU q FATTORI}\quad \frac{\lambda_1+\dots+\lambda_q}{\sum_\alpha\lambda_\alpha}$$

• Analisi Generale



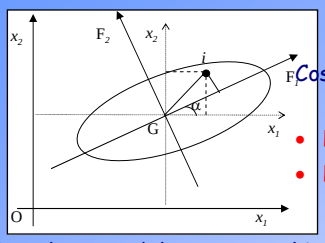
• *p* dimensioni

• necessità di *più fattori di sintesi*

• La specificazione dei vincoli → **Ortogonalità dei fattori**

• La scelta del numero di fattori → **Scree-test, Eigenvalue-one, % variabilità spiegata, ...**

I PASSI DELL'ACP



$F_1 = x_1 \operatorname{sen}\alpha + x_2 \cos\alpha$

Coseni direttori dell'asse F_1

- L'asse F_1 è a distanza minima dai punti *i* della nube
- La loro immagine (proiezione ψ_{i1}) su F_1 è max

① Traslazione del sistema nel baricentro G (centratura var.)

② Rotazione nella direzione di massimo allungamento della nube

Ricerca dell'asse F1 tale che $\sum_{i=1}^{n} n_i \psi_{i1}^2 = max$

Autovalore $\lambda_1 = \sum n_i \psi_{i1}^2$ inerzia (varianza) nella direzione di max allungamento della nube

Autovettore $\mathbf{u}_1 = [\mathbf{u}_1^1 \ \mathbf{u}_a^2]$ individua la giacitura dell'asse F1 (coseni direttori dell'asse $\|\mathbf{u}_1\| = 1$ di max allungamento

Componenti principali: $\psi_{i1} = x_i^1 \mathbf{u}_1^1 + x_i^2 \mathbf{u}_1^2 \quad i = 1, \ldots, n$

$\mathbf{VMu}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \quad \mathbf{u}'_\alpha \mathbf{u}_\alpha = 1, \ \mathbf{u}'_\alpha \mathbf{u}_{\alpha'} = 0 \quad (\alpha \neq \alpha')$

Equazione spettrale:

③ Ricerca di una direttrice F_2 di max allungamento ortogonale a F_1

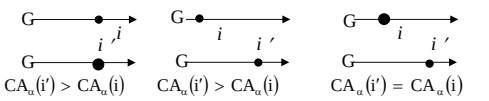
Ricerca dell'asse F_2 tale che $\boxed{\sum_{i=1}^{n} n_i \psi_{i2}^2 = \max}$

Calcolo di λ_2 e $\mathbf{u}_2 = [\mathbf{u}_1^2 \;\; \mathbf{u}_a^2]$

④ Qualità della rappresentazione su F_1 $\boxed{\tau_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\sum n_i \psi_{i1}^2}{\mathrm{var}(x_1) + \mathrm{var}(x_2)}}$

CONTRIBUTI ASSOLUTI

$\underset{(u)}{\mathrm{Max}} \sum_i \{f_{i.} \;\; \psi_{i\alpha}^2\} = \lambda_\alpha$ $\boxed{CA_\alpha(i) = f_{i.} \;\; \psi_{i\alpha}^2 / \lambda_\alpha}$ $\left(\sum_\alpha CA_\alpha(i) = 1\right)$



$CA_\alpha(i') > CA_\alpha(i)$ $CA_\alpha(i') > CA_\alpha(i)$ $CA_\alpha(i') = CA_\alpha(i)$

$f_{i'} > f_i$ $\psi_{i'\alpha} > \psi_{i\alpha}$ $f_{i'} \;\; \psi_{i\alpha}^2 > f_i \;\; \psi_{i'\alpha}^2$

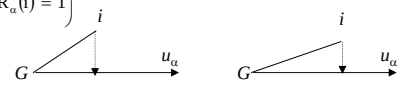
Evidenziano il ruolo di un elemento nella spiegazione del fattore α- esimo

CONTRIBUTI RELATIVI

$\boxed{CR_\alpha(i) = d_\alpha^2(i,G) / d^2(i,G)}$

$d_\alpha^2(i,G) = \psi_{i\alpha}^2$ $d^2(i,G) = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)$

$\left(\sum_\alpha CR_\alpha(i) = 1\right)$



i è mal rappresentato *i* è ben rappresentato

- Misura la qualità della rappresentazione di un punto su un asse
- Per analizzare le prossimità tra i punti ci si interesserà di quelli con forte CR

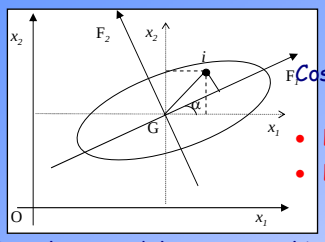
• Analisi in Componenti Principali (ACP)

Tabella individui × variabili:

$$X = \begin{array}{c|cccc} & x_1 & x_2 & \cdots & x_p \\ \hline i_1 & 21{,}1 & 3{,}2 & \vdots & 12{,}6 \\ i_2 & 15{,}6 & 8{,}4 & & 17{,}2 \\ \vdots & \cdots\cdots & & x_{ij} & \cdots\cdots \\ \vdots & & & \vdots & \\ i_n & 16{,}4 & 7{,}2 & & 21{,}3 \end{array}$$

- Le righe rappresentano gli individui e sono in genere osservazioni, oggetti, unità statistiche;
- Le colonne rappresentano le variabili definite da valori numerici continui;
- Gli angoli tra punti-variabile possono essere interpretati in termini di correlazioni tra le stesse;
- Le prossimità tra punti-individuo possono essere interpretate in termini di analogie di comportamento rispetto al fenomeno osservato.

I PASSI DELL'ACP



$F_1 = x_1 \operatorname{sen}\alpha + x_2 \cos\alpha$

Coseni direttori dell'asse F_1

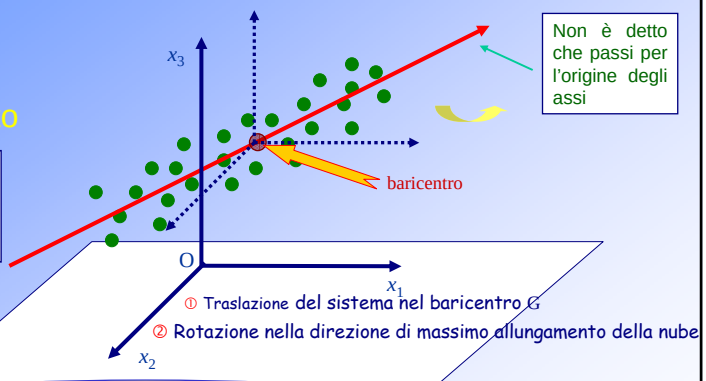
- L'asse F_1 è a distanza minima dai punti *i* della nube
- La loro immagine (proiezione ψ_{i1}) su F_1 è max

① Traslazione del sistema nel baricentro G (centratura var.)

② Rotazione nella direzione di massimo allungamento della nube

- **Analisi in Componenti Principali (ACP)**
  - Spazio R^p
  - Analisi nello spazio centrato

Ad ogni variabile si sottrae la propria media: $x^*_{ij} = x_{ij} - \bar{x}_j$



Non è detto che passi per l'origine degli assi

① Traslazione del sistema nel baricentro G

② Rotazione nella direzione di massimo allungamento della nube

L'ACP è un'Analisi generale su variabili centrate

$$\underset{(H)}{\text{Max}} \left\{ \sum_i \sum_{i'} d^2_H(i,i') \right\} = \underset{(H)}{\text{Max}} \left\{ \sum_i d^2_H(i,G) \right\} \Rightarrow x^*_{ij} = \frac{1}{\sqrt{n}} \left(x_{ij} - \bar{x}_j \right)$$

DIAGONALIZZARE X* ' X* = [covarianza]

• Esempio ACP: i consumi alimentari

Matrice dei dati: 16 Paesi, 10 variabili continue

	Cereali	Riso	Patate	Zucchero	Verdure	Vino	Carne	Latte	Burro	Uova
Belgio	72,2	4,2	98,8	40,4	103,2	20,9	102,0	80,0	7,7	14,2
Danimarca	70,5	2,2	57,0	39,5	50,0	22,0	105,8	145,2	4,1	14,3
Germania	71,3	2,3	74,1	37,1	83,1	22,8	97,2	90,7	6,9	14,8
Grecia	109,8	5,4	90,0	30,0	229,5	25,3	77,1	63,1	0,9	11,3
Spagna	71,4	5,8	107,8	26,8	191,7	43,0	102,1	98,4	0,6	15,3
Francia	73,0	4,3	78,2	34,1	95,0	64,5	110,5	98,9	8,9	15,0
Irlanda	93,4	3,2	151,5	34,8	55,0	3,9	105,0	185,9	3,4	11,4
Italia	110,2	4,8	38,6	27,9	181,9	61,6	88,0	65,0	2,4	11,1
Olanda	54,6	5,0	86,7	39,7	99,0	14,0	89,4	136,2	5,4	10,7
Portogallo	86,0	5,7	106,6	29,4	100,0	57,0	75,5	96,0	1,5	7,7
RegnoUnito	74,3	4,5	94,1	39,8	60,0	10,4	74,4	129,3	3,2	10,8
Austria	68,7	4,2	62,6	37,1	81,9	34,3	93,4	121,3	4,3	13,4
Finlandia	70,1	5,4	61,6	35,7	52,6	10,2	65,0	208,4	5,8	10,9
Islanda	79,7	1,9	50,2	54,9	50,0	6,2	71,7	205,6	4,6	11,3
Norvegia	76,9	3,5	73,2	37,3	48,3	6,6	54,9	176,5	2,1	11,3
Svezia	69,3	4,3	70,0	37,5	48,5	12,3	60,5	154,1	5,7	12,9

• Esempio ACP: i consumi alimentari

Statistiche delle variabili continue

```
EFFETTIVI TOTALI :      16              PESO TOTALE :      16.00
+---------------------------+-----------------+---------------+
|IDEN – ETICH.   EFF.  PESO  |  MEDIA     SQM  |  MIN     MAX  |
+---------------------------+-----------------+---------------+
|CERE - cereali   16   16.00 |   78.21   14.45 | 54.60  110.20 |
|RISO - riso      16   16.00 |    4.17    1.20 |  1.90    5.80 |
|PATA - patate    16   16.00 |   81.31   26.48 | 38.60  151.50 |
|ZUCC - zucchero  16   16.00 |   36.38    6.40 | 26.80   54.90 |
|VERD - verdure   16   16.00 |   95.61   55.01 | 48.30  229.50 |
|VINO - vino      16   16.00 |   25.94   19.69 |  3.90   64.50 |
|CARN - carne     16   16.00 |   85.78   17.05 | 54.90  110.50 |
|LATT - latte     16   16.00 |  128.41   46.03 | 63.10  208.40 |
|BURR - burro     16   16.00 |    4.22    2.35 |  0.60    8.90 |
|UOVA - uova      16   16.00 |   12.27    2.01 |  7.70   15.30 |
+---------------------------+-----------------+---------------+
```

Affinchè $d^2(i,i')$ sia indipendente dalle unità di misura di j si ha:

$$d^2(i,i') = \sum_{j=1}^{p} \left\{ \frac{1}{\sqrt{n}} \cdot \frac{r_{ij} - \bar{r}_j}{s_j} - \frac{1}{\sqrt{n}} \cdot \frac{r_{i'j} - \bar{r}_j}{s_j} \right\}^2$$

DIAGONALIZZARE X*' X* = [correlazione] (Analisi Generale di **X**)

$$\sum_i \hat{\psi}_{i\alpha} = 0$$

• Esempio ACP: i consumi alimentari

Matrice di correlazione

	Cere	Riso	Pata	Zucc	Verd	Vino	Carn	Latt	Burr	Uova
Cere	1,00									
Riso	0,13	1,00								
Pata	0,06	0,23	1,00							
Zucc	-0,41	-0,69	-0,28	1,00						
Verd	0,56	0,57	0,07	-0,64	1,00					
Vino	0,29	0,42	-0,13	-0,62	0,54	1,00				
Carn	-0,07	-0,15	0,29	-0,19	0,22	0,39	1,00			
Latt	-0,34	-0,39	-0,04	0,58	-0,75	-0,69	-0,41	1,00		
Burr	-0,52	-0,34	-0,19	0,43	-0,46	-0,06	0,29	0,10	1,00	
Uova	-0,34	-0,31	-0,10	0,02	0,07	0,11	0,60	-0,22	0,45	1,00

• Esempio ACP: i consumi alimentari

Autovettori

	1	2	3	4	5
Cere – cereali	0.30	0.24	0.18	-0.61	-0.27
Riso – riso	0.35	0.18	-0.13	0.62	0.09
Pata – patate	0.11	0.05	-0.86	-0.10	-0.15
Zucc – zucchero	-0.44	-0.01	0.15	-0.18	-0.05
Verd – verdure	0.45	-0.03	0.11	-0.13	0.37
Vino – vino	0.37	-0.23	0.27	0.17	-0.46
Carn – carne	0.12	-0.54	-0.29	-0.26	-0.20
Latt – latte	-0.40	0.26	-0.14	0.02	0.02
Burr – burro	-0.25	-0.42	0.08	0.28	-0.42
Uova – uova	-0.04	-0.57	-0.01	-0.09	0.58

• Esempio ACP: i consumi alimentari

Coordinate sui nuovi assi

• Es.: Belgio sul primo fattore

Valori iniziali

Cere	Riso	Pata	Zucc	Verd	Vino	Carn	Latt	Burr	Uova
72,2	4,2	98,8	40,4	103,2	20,9	102,0	80,0	7,7	14,2

$$\begin{bmatrix} -0{,}42 & 0{,}03 & 0{,}66 & 0{,}63 & 0{,}14 & -0{,}26 & 0{,}95 & -1{,}05 & 1{,}48 & 0{,}96 \end{bmatrix} \times \begin{pmatrix} 0{,}30 \\ 0{,}35 \\ 0{,}11 \\ -0{,}44 \\ 0{,}45 \\ 0{,}37 \\ 0{,}12 \\ -0{,}40 \\ -0{,}25 \\ -0{,}04 \end{pmatrix}$$

Primo autovettore

Dati standardizzati

$= (-0{,}42 \times 0{,}30) + (0{,}03 \times 0{,}35) + (0{,}66 \times 0{,}11) + \ldots + (0{,}96 \times -0{,}04)$

$= -0{,}23$

Coordinata del BELGIO sul primo fattore

• Esempio ACP: i consumi alimentari

Coordinate sui nuovi assi

	1	2	3	4	5
BELGIO	-0.23	-1.96	-0.62	0.10	-0.12
DANIMARCA	-1.52	-1.48	0.43	-1.00	0.28
GERMANIA	-0.85	-2.12	0.30	-0.64	0.12
GRECIA	3.45	1.43	0.32	-1.14	0.78
SPAGNA	2.92	-0.95	-1.23	0.39	1.72
FRANCIA	0.68	-3.08	0.32	0.77	-1.10
IRLANDA	-0.58	0.62	-2.93	-1.78	-0.81
ITALIA	3.42	0.31	2.27	-0.75	-0.57
OLANDA	-0.79	0.04	-0.67	1.36	0.05
PORTOGALLO	2.40	1.96	-0.42	0.83	-1.55
REGNO UNITO	-0.70	1.17	-0.54	0.25	0.01
AUSTRIA	-0.18	-0.89	0.48	0.42	0.21
FINLANDIA	-1.48	1.43	0.26	1.63	0.05
ISLANDA	-3.58	0.95	1.35	-1.36	-0.14
NORVEGIA	-1.46	2.02	0.29	-0.07	0.63
SVEZIA	-1.49	0.54	0.39	0.99	0.46

• Esempio ACP: i consumi alimentari

La variabilità sui nuovi fattori: gli autovalori

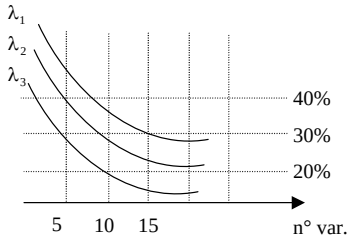
Num	Autov.	%	% cum.	Istogramma
1	3.8920	38.92	38.92	••
2	2.2966	22.97	61.89	••••••••••••••••••••••••
3	1.2126	12.13	74.01	••••••••••••
4	0.9672	9.67	83.68	••••••••••
5	0.5595	5.59	89.28	••••••
6	0.4064	4.06	93.34	••••
7	0.3050	3.05	96.39	•••
8	0.2148	2.15	98.54	••
9	0.1091	1.09	99.63	•
10	0.0367	0.37	100.00	•

Numero degli assi (rappresentazione ottimale)

- percentuale varianza $\tau_h = \lambda_1 + \ldots + \lambda_h / \lambda_1 + \ldots + \lambda_p \geq 0.75$
- eigenvalue one $\lambda_1 + \ldots + \lambda_h > 1$
- scree test:

  (max caduta istogr. autovalori)

- abachi



ACP ranghi ; ACB

num. autovalori tale che:

$\tau_h > \tau_{tabulati}$

• Esempio ACP: i consumi alimentari

Le coordinate nel vecchio e nel nuovo spazio

	Variabili iniziali										Componenti Principali				
	Ce	Ri	Pa	Zu	Ve	Vi	Ca	La	Bu	Uo	1	2	3	4	5
Belg	72,2	4,2	98,8	40,4	103,2	20,9	102,0	80,0	7,7	14,2	-0,23	-1,97	-0,63	0,09	-0,13
Dani	70,5	2,2	57,0	39,5	50,0	22,0	105,8	145,2	4,1	14,3	-1,53	-1,49	0,43	-1,01	0,27
Germ	71,3	2,3	74,1	37,1	83,1	22,8	97,2	90,7	6,9	14,8	-0,86	-2,13	0,30	-0,65	0,11
Grec	109,8	5,4	90,0	30,0	229,5	25,3	77,1	63,1	0,9	11,3	3,45	1,44	0,33	-1,12	0,80
Spag	71,4	5,8	107,8	26,8	191,7	43,0	102,1	98,4	0,6	15,3	2,91	-0,95	-1,24	0,39	1,73
Fran	73,0	4,3	78,2	34,1	95,0	64,5	110,5	98,9	8,9	15,0	0,67	-3,08	0,31	0,74	-1,12
Irla	93,4	3,2	151,5	34,8	55,0	3,9	105,0	185,9	3,4	11,4	-0,57	0,63	-2,93	-1,78	-0,81
Ital	110,2	4,8	38,6	27,9	181,9	61,6	88,0	65,0	2,4	11,1	3,41	0,32	2,28	-0,75	-0,56
Olan	54,6	5,0	86,7	39,7	99,0	14,0	89,4	136,2	5,4	10,7	-0,79	0,04	-0,68	1,36	0,04
Port	86,0	5,7	106,6	29,4	100,0	57,0	75,5	96,0	1,5	7,7	2,40	1,97	-0,41	0,84	-1,54
GrBr	74,3	4,5	94,1	39,8	60,0	10,4	74,4	129,3	3,2	10,8	-0,70	1,17	-0,54	0,26	0,01
Aust	68,7	4,2	62,6	37,1	81,9	34,3	93,4	121,3	4,3	13,4	-0,19	-0,89	0,48	0,41	0,21
Finl	70,1	5,4	61,6	35,7	52,6	10,2	65,0	208,4	5,8	10,9	-1,48	1,44	0,26	1,64	0,04
Isla	79,7	1,9	50,2	54,9	50,0	6,2	71,7	205,6	4,6	11,3	-3,57	0,94	1,37	-1,36	-0,16
Norv	76,9	3,5	73,2	37,3	48,3	6,6	54,9	176,5	2,1	11,3	-1,46	2,02	0,29	-0,06	0,63
Svez	69,3	4,3	70,0	37,5	48,5	12,3	60,5	154,1	5,7	12,9	-1,48	0,54	0,39	0,99	0,46

• Esempio ACP: i consumi alimentari

L'interpretazione dei fattori: correlazioni tra variabili iniziali e componenti principali

	Ce	Ri	Pa	Zu	Ve	Vi	Ca	La	Bu	Uo	1	2	3	4	5
Ce	1,00														
Ri	0,13	1,00													
Pa	0,06	0,23	1,00												
Zu	-0,41	-0,69	-0,28	1,00											
Ve	0,56	0,57	0,07	-0,64	1,00										
Vi	0,29	0,42	-0,13	-0,62	0,54	1,00									
Ca	-0,07	-0,15	0,29	-0,19	0,22	0,39	1,00								
La	-0,34	-0,39	-0,04	0,58	-0,75	-0,69	-0,41	1,00							
Bu	-0,52	-0,34	-0,19	0,43	-0,46	-0,06	0,29	0,10	1,00						
Uo	-0,34	-0,31	-0,10	0,02	0,07	0,11	0,60	-0,22	0,45	1,00					
1	0,59	0,69	0,21	-0,87	0,89	0,73	0,24	-0,79	-0,50	-0,07	1,00				
2	0,36	0,27	0,07	-0,01	-0,05	-0,35	-0,81	0,39	-0,64	-0,86	0,00	1,00			
3	0,19	-0,14	-0,95	0,16	0,12	0,30	-0,32	-0,16	0,09	-0,01	0,00	0,00	1,00		
4	-0,60	0,61	-0,10	-0,18	-0,13	0,17	-0,25	0,02	0,28	-0,09	0,00	0,00	0,00	1,00	
5	-0,20	0,07	-0,11	-0,04	0,28	-0,34	-0,15	0,02	-0,31	0,43	0,00	0,00	0,00	0,00	1,00

• Esempio ACP: i consumi alimentari

Le coordinate delle variabili

		1	2	3	4	5
CERE	- cereali	0.59	0.36	0.19	-0.60	-0.20
RISO	- riso	0.69	0.27	-0.14	0.61	0.07
PATA	- patate	0.21	0.07	-0.95	-0.10	-0.11
ZUCC	- zucchero	-0.87	-0.01	0.16	-0.18	-0.04
VERD	- verdure	0.89	-0.05	0.12	-0.13	0.28
VINO	- vino	0.73	-0.35	0.30	0.17	-0.34
CARN	- carne	0.24	-0.81	-0.32	-0.25	-0.15
LATT	- latte	-0.79	0.39	-0.16	0.02	0.02
BURR	- burro	-0.50	-0.64	0.09	0.28	-0.31
UOVA	- uova	-0.07	-0.86	-0.01	-0.09	0.43

• Esempio ACP: i consumi alimentari (note)

DISTANZA TRA VARIABILI

$$d^2(0,\ j) = \sum_i x_{ij}^2 = \sum_i \left(\frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}} \right)^2 = \frac{1}{s_j^2} \cdot \frac{1}{n} \cdot \sum_i \left(r_{ij} - \bar{r}_j \right)^2 = 1$$

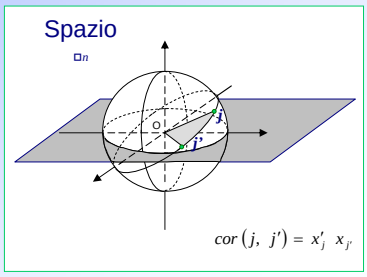
$$d^2(j,\ j') = \sum_i \left(x_{ij} - x_{ij'} \right)^2 = \sum_i x_{ij}^2 + \sum_i x_{ij'}^2 - 2 \sum_i x_{ij}\ x_{ij'}$$

$$d^2(j,\ j') = 2\left[1 - cor(j,\ j')\right]$$

$cor(j,\ j') \approx 1$ punti vicini

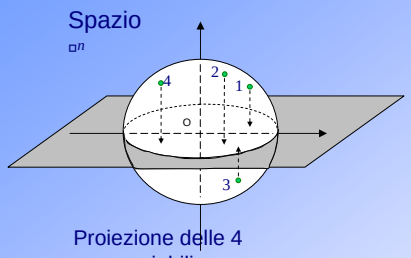
$cor(j,\ j') \approx 0$ punti a distanza media

$cor(j,\ j') \approx -1$ punti molti distanti

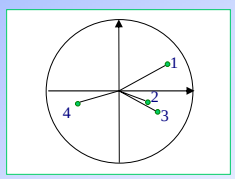


• Esempio ACP: i consumi alimentari (note)

Spazio



Proiezione delle 4 variabili



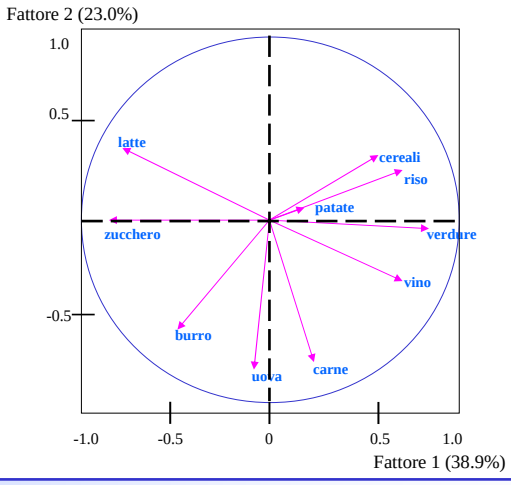
Piano Fattoriale

- La nube non è centrata su 0
- Un asse è una VARIABILE ARTIFICIALE combinazione lineare di *p* variabili
- $\hat{\varphi} = \mathbf{X}' \mathbf{v}$ ⇒ {ascissa di "*j*" su $\mathbf{v}_\alpha$} $= \hat{\varphi}_{j\alpha} = \mathbf{x}'_j \mathbf{v}_\alpha = cor\left(\text{"j"}, \text{"}\alpha\text{"}\right)$

  (utile per l'interpretazione degli assi)
- FATTORE TAGLIA: se le maggior parte delle variabili sono tra loro correlate positivamente questo equivale a dire che sono simultanemante forti o deboli

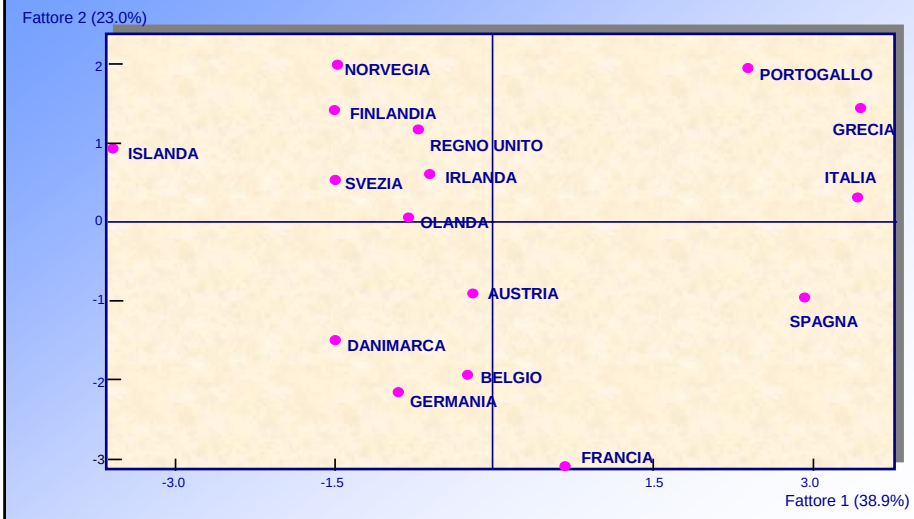
• Esempio ACP: i consumi alimentari

Le rappresentazioni grafiche: il Cerchio delle correlazioni



• Esempio ACP: i consumi alimentari

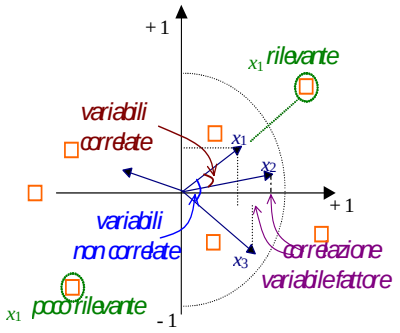
Le rappresentazioni grafiche: il Piano degli individui



Rappresentazione Simultanea

Analisi in Componenti Principali

→ lettura angoli

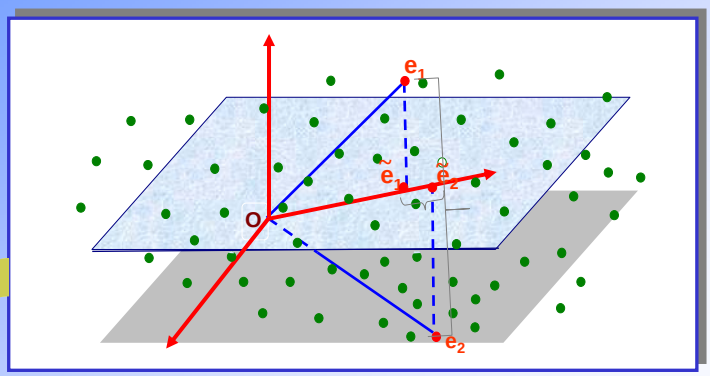


• L'interpretazione dei fattori

1) $$\lambda_\alpha = \sum_{I=1}^{n} \tilde{e}_{i\alpha}^2 \quad ; \quad CA_{i\alpha} = \frac{\tilde{e}_{i\alpha}^2}{\lambda_\alpha}$$

Un **autovalore** è una misura della capacità esplicativa dei singoli fattori. Il **contributo assoluto** esprime l'apporto informativo di ogni unità alla variabilità dell'asse.

2)

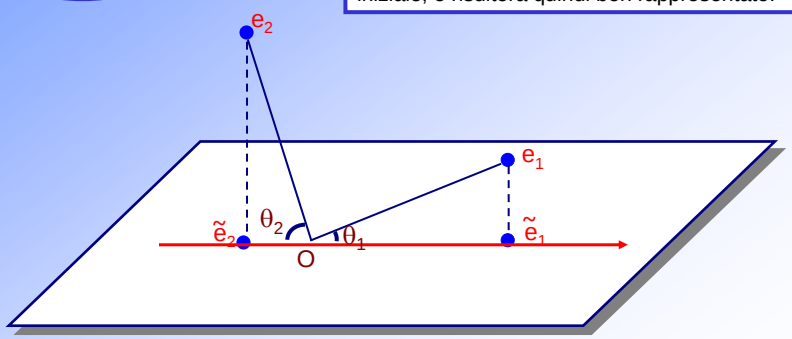


Due punti-unità possono trovarsi vicini nello spazio di proiezione non perché lo siano nello spazio iniziale di riferimento ma solo per gli ***effetti distorsivi*** dell'operazione di proiezione

• Una misura della qualità della rappresentazione

$$CR_{i\alpha} = \cos^2_{i\alpha} = \frac{\|\tilde{e}_{i\alpha}\|^2}{\|e_i\|^2}$$

Il **coseno al quadrato** è una misura della qualità della rappresentazione dei punti-unità sul sottospazio generato dai fattori scelti: quanto più risulta prossimo ad **1** tanto più il punto avrà conservato, in proiezione, la distanza dall'origine che aveva nello spazio iniziale, e risulterà quindi ben rappresentato.



• Esempio i consumi alimentari

	Contributi assoluti delle unità all'inerzia degli assi principali					Coseni al quadrato degli angoli tra i vettori e gli assi principali				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Belgio	0,30	12,30	0,70	0,20	11,40	0,02	0,69	0,02	0,01	0,14
Danimarca	5,40	4,20	2,20	5,40	9,20	0,41	0,20	0,06	0,11	0,10
Germania	2,30	11,60	2,10	1,10	0,30	0,19	0,59	0,06	0,02	0,00
Grecia	26,70	1,20	5,00	2,80	0,30	0,85	0,02	0,05	0,02	0,00
Spagna	13,10	6,40	4,70	1,40	45,30	0,51	0,15	0,06	0,01	0,25
Francia	0,30	20,10	0,00	2,30	7,90	0,02	0,81	0,00	0,04	0,07
Irlanda	0,00	0,10	33,40	38,30	0,80	0,00	0,00	0,49	0,46	0,01
Italia	15,40	0,20	28,80	0,90	2,70	0,57	0,00	0,35	0,01	0,01
Olanda	0,70	0,00	4,30	10,50	1,60	0,09	0,00	0,18	0,36	0,03
Portogallo	8,90	10,80	6,00	0,40	8,20	0,43	0,33	0,10	0,01	0,06
Gr. Bret.	0,20	3,40	2,00	0,10	0,40	0,04	0,52	0,16	0,01	0,01
Austria	0,40	1,40	0,70	1,50	2,40	0,15	0,30	0,08	0,15	0,12
Finlandia	2,30	7,60	0,30	16,50	0,00	0,16	0,33	0,01	0,31	0,00
Islanda	19,60	6,70	9,50	10,60	0,00	0,61	0,13	0,10	0,09	0,00
Norvegia	1,50	12,70	0,20	0,00	9,10	0,13	0,66	0,00	0,00	0,11
Svezia	2,90	1,50	0,10	7,90	0,30	0,37	0,12	0,01	0,27	0,01

• ACP: Un esempio “di corsa”

Paese	100 m.	200 m.	400 m.	800m.	1500 m.	5 km.	10 km.	Maratona
Austria	10,44	20,81	46,82	1,79	3,60	13,26	27,72	135,90
Belgio	10,34	20,68	45,04	1,73	3,60	13,22	27,45	129,95
Cecoslov	10,35	20,65	45,64	1,76	3,58	13,42	28,19	134,32
Danimarca	10,56	20,52	45,89	1,78	3,61	13,50	28,11	130,78
Finlandia	10,43	20,69	45,49	1,74	3,61	13,27	27,52	130,87
Francia	10,11	20,38	45,28	1,73	3,57	13,34	27,97	132,30
Germania	10,12	20,33	44,50	1,73	3,53	13,17	27,42	129,92
GB	10,11	20,21	44,93	1,70	3,51	13,01	27,51	129,13
Grecia	10,22	20,71	46,56	1,78	3,64	14,59	28,45	134,60
Irlanda	10,61	20,96	46,30	1,79	3,56	13,32	27,81	132,35
Italia	10,01	19,72	45,26	1,73	3,60	13,23	27,52	131,08
Lussemb	10,35	20,77	47,40	1,82	3,67	13,64	29,08	141,27
Norvegia	10,55	21,16	46,71	1,76	3,62	13,34	27,69	131,48
Olanda	10,52	20,95	45,10	1,74	3,62	13,36	27,61	129,02
Polonia	10,16	20,24	45,36	1,76	3,60	13,29	27,89	131,58
Portogallo	10,53	21,17	46,70	1,79	3,62	13,13	27,38	128,65
Romania	10,41	20,98	45,87	1,76	3,64	13,25	27,67	132,50
Russia	10,07	20,00	44,60	1,75	3,59	13,20	27,53	130,55
Spagna	10,42	20,77	45,98	1,76	3,55	13,31	27,73	131,57
Svezia	10,25	20,61	45,63	1,77	3,61	13,29	27,94	130,63
Svizzera	10,37	20,46	45,78	1,78	3,55	13,22	27,91	131,20
Ungheria	10,26	20,62	46,02	1,77	3,62	13,49	28,44	132,58
USA	9,93	19,75	43,86	1,73	3,53	13,20	27,43	128,22

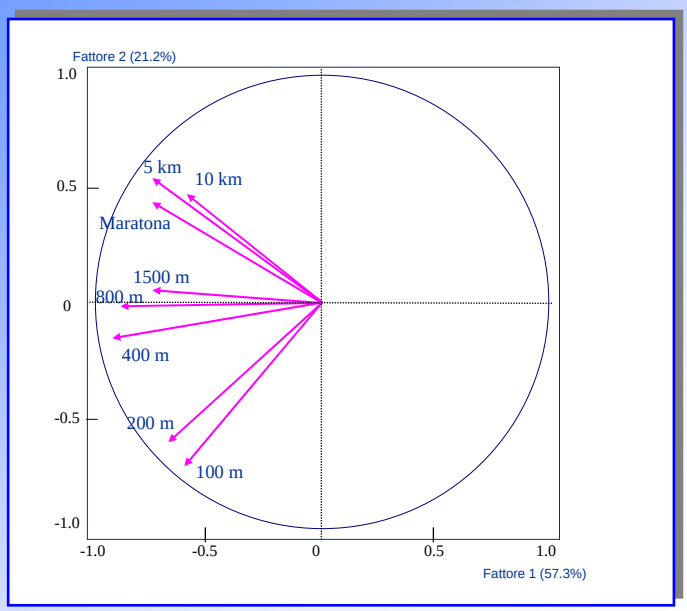
• Matrice di correlazione

	100 m.	200 m.	400 m.	800m.	1500 m.	5 km.	10 km.	Mara-tona
100 m.	1,00							
200m.	0,88	1,00						
400m.	0,65	0,72	1,00					
800m.	0,55	0,53	0,82	1,00				
1500 m.	0,38	0,48	0,62	0,57	1,00			
5 km.	0,06	0,19	0,42	0,41	0,49	1,00		
10 km.	0,11	0,19	0,58	0,63	0,50	0,66	1,00	
Maratona	0,16	0,26	0,69	0,65	0,51	0,49	0,81	1,00

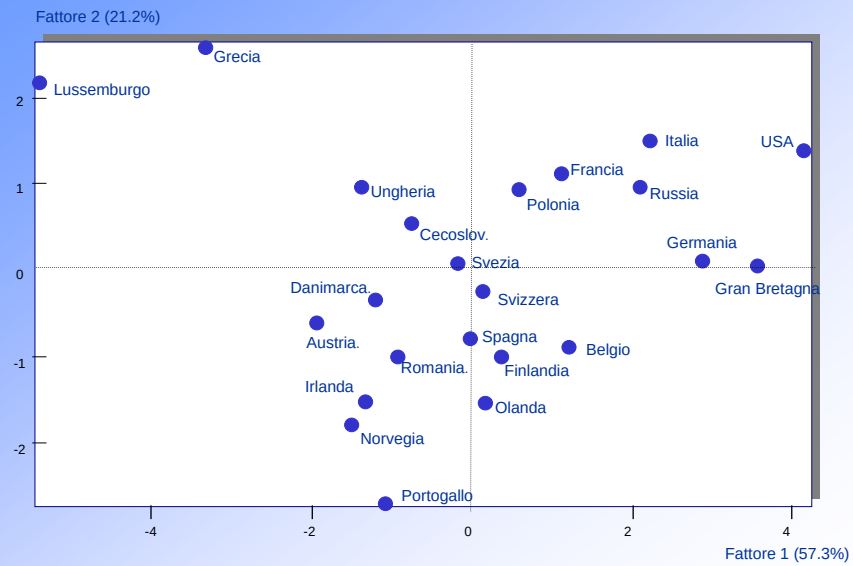
• Autovalori

N.	Valore	%	% cum.	Istogramma
1	4.5822	57.3	57.3	••
2	1.6967	21.2	78.5	•••••••••••••••••••••
3	0.6288	7.9	86.4	••••••••
4	0.4429	5.5	91.9	••••••
5	0.2763	3.5	95.3	••••
6	0.1884	2.4	97.7	••
7	0.1015	1.3	99.0	•
8	0.0832	1.0	100.0	•

• Il cerchio delle correlazioni



• La rappresentazione delle unità



Analisi delle Corrispondenze (ACB)

Obiettivo: **Studiare la struttura dell'associazione tra due o più variabili qualitative. L'AC presuppone che i legami tra le variabili siano di tipo simmetrico**

- L'AC permette di decomporre una tabella a due o più entrate in una serie di fattori ciascuno dei quali rappresenta un aspetto latente del tipo di associazione presente nei dati
- La rappresentazione in forma grafica dei fattori consente una interpretazione semplice della struttura dell'associazione e permette di evidenziare aspetti non direttamente rilevabili alla lettura della tabella

I dati dell'AC

Considerati due insiemi finiti I e J, non necessariamente muniti di struttura d'ordine, per convenzione indicati

con {I: i∈ 1…I} e {J: j∈ 1…J},

una applicazione f definisce una corrispondenza tra I e J

$$\forall (i,j) \in I \times J \underset{f}{\rightarrow} x_{ij} \in \Re^{+}$$

Esempi

Medicina

I I ⇒ insieme dei [illegible]iti
J J ⇒ insieme delle malattie
$x X_{ij}$ ⇒ n° delle volte (frequenze) che il trattamento *i* è stato somministrato per la malattia *j*

1 ……..j ……..J

X → 1 . . . i . . . I | x_{ij}

Tabelle di
-Contingenza
- Misure omogenee
-Dati booleani

Economia

I I ⇒ insieme dei settori venditori
J J ⇒ insieme dei settori acquirenti
$X X_{ij}$ ⇒ valore degli scambi tra i settori *i* e *j* (per le tavole input-output I≡J)

Ecologia

I I ⇒ insieme delle stazioni di rilevazione
J J ⇒ insieme delle specie
$X X_{ij}$ ⇒ 1 (0) presenza (assenza) della specie *j* nella stazione *i*

ma anche
- Codifica disgiuntiva
- Dati testuali

AC: un esempio

Prodotti

Consumatori

N =

	P1	P2	P3	P4	Totale
A	69	37	7	5	118
B	148	45	14	22	229
C	170	65	12	29	276
D	159	57	12	28	256
E	122	26	6	18	172
F	106	21	5	23	155
G	40	7	1	14	62
Totale	814	258	57	139	1268

$$n_{i.} = \sum_{j=1}^{c} n_{ij}$$

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

$$n = \sum_{i,j} n_{ij}$$

F ➡ matrice delle frequenze relative

$$f_{ij} = \frac{n_{ij}}{n}$$

F =

	P1	P2	P3	P4	Totale
A	0,054	0,029	0,006	0,004	0,0931
B	0,117	0,035	0,011	0,017	0,1806
C	0,134	0,051	0,009	0,023	0,2177
D	0,125	0,045	0,009	0,022	0,2019
E	0,096	0,021	0,005	0,014	0,1356
F	0,084	0,017	0,004	0,018	0,1222
G	0,032	0,006	0,001	0,011	0,0489
Totale	0,642	0,2035	0,045	0,1096	1,000

$$f_{.j} = \sum_{i=1}^{r} f_{ij}$$

$$f_{i.} = \sum_{j=1}^{c} f_{ij}$$

$$\sum_{i,j} f_{ij} = 1$$

•Il test di indipendenza

$$X = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

$$\hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

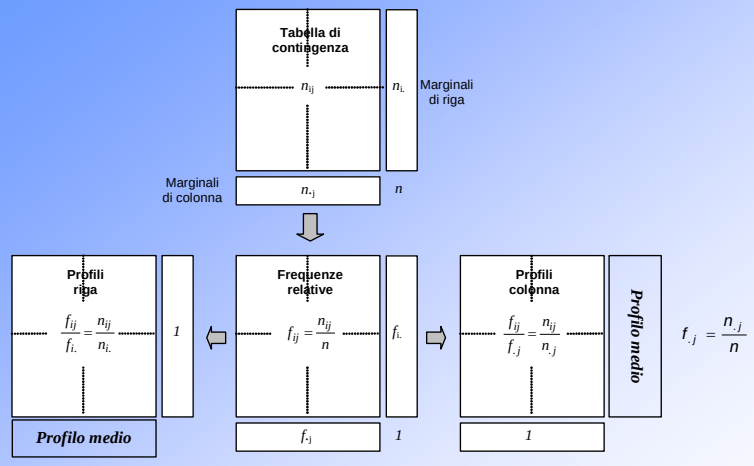
Sotto l'ipotesi di indipendenza

$$X = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

Test di indipendenza:

VALORE DEL CHI-QUADRATO	=	39.054
GRADI DI LIBERTA'	=	18.000
P-VALUE	=	.004

AC: le tabelle



$$f_{i.} = \frac{n_{i.}}{n}$$

PROFILO RIGA

$$\frac{f_{ij}}{f_{i.}} = \frac{n_{ij}}{n} \cdot \frac{n}{n_{i.}} = \frac{n_{ij}}{n_{i.}}$$

PROFILO COLONNA

$$\frac{f_{ij}}{f_{.j}} = \frac{n_{ij}}{n} \cdot \frac{n}{n_{.j}} = \frac{n_{ij}}{n_{.j}}$$

AC: un esempio

Profili riga

	P1	P2	P3	P4	*Totale*
A	58,5	31,4	5,9	4,2	100,0
B	64,6	19,7	6,1	9,6	100,0
C	61,6	23,6	4,3	10,5	100,0
D	62,1	22,3	4,7	10,9	100,0
E	70,9	15,1	3,5	10,5	100,0
F	68,4	13,5	3,2	14,8	100,0
G	64,5	11,3	1,6	22,6	100,0
Profilo medio	*64,2*	*20,3*	*4,5*	*11,0*	100,0

Comportamento atteso

Profili colonna

	P1	P2	P3	P4	*Profilo medio*
A	8,5	14,3	12,3	3,6	9,3
B	18,2	17,4	24,6	15,8	18,1
C	20,9	25,2	21,1	20,9	21,8
D	19,5	22,1	21,1	20,1	20,2
E	15,0	10,1	10,5	12,9	13,6
F	13,0	8,1	8,8	16,5	12,2
G	4,9	2,7	1,8	10,1	4,9
Totale	100,0	100,0	100,0	100,0	*100,0*

Comportamento atteso

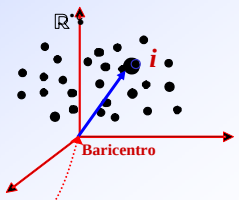
AC: Interpretazione geometrica

- Ogni **profilo** riga (o colonna) può essere considerato come un **vettore** in uno spazio multidimensionale

- A differenza dell'ACP, dove ciascun punto-individuo è caratterizzato da un peso costante e pari ad 1/n, nell'AC **ciascun punto è munito di una massa** pari al rapporto tra il rispettivo marginale ed il totale generale della tabella, così che l'informazione relativa delle singole righe e colonne sia comunque conservata

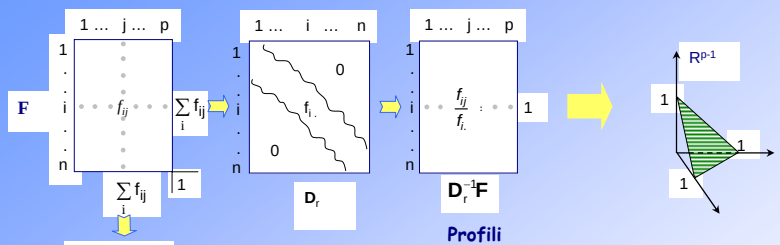
	58.5	31.4	5.9	4.2	×	0.093	+	
	64.6	19.7	6.1	9.6	×	0.181	+	
	61.6	23.6	4.3	10.5	×	0.218	+	
Profili → *i*	62.1	22.3	4.7	10.9	×	0.202	+	
	70.9	15.1	3.5	10.5	×	0.136	+	
	68.4	13.5	3.2	14.8	×	0.122	+	
	64.5	11.3	1.6	22.6	×	0.049	=	
Profilo medio →	***64.2***	***20.3***	***4.5***	***11.0***				

Masse → 118 / 1268

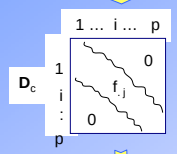


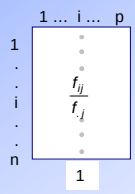
La j-esima componente del profilo medio può essere ottenuta come media delle j-esime componenti degli r profili riga, ciascuna ponderata con la corrispondente massa

Spazi dei profili

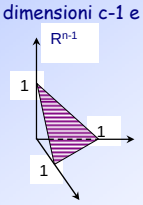


Profili Colonna





Profili Colonna



- Il vincolo di somma ad 1 per riga o colonna porta alla perdita di una dimensione nei rispettivi spazi di riferimento
- Rappresentazione dei profili riga e colonna in spazi di dimensioni c-1 e r-1

- Nell'ACP l'operazione di centratura (standardizzazione) viene effettuata sulle colonne della matrice di partenza
- Nell'AC le stesse trasformazioni vengono effettuate nei due spazi: le variabili giocano ruoli simmetrici

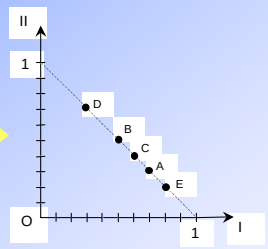
•Autovalore Banale

Tabella di contingenza

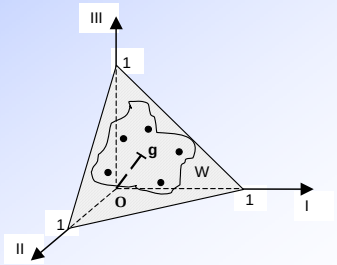
	I	II	Tot.
A	7	3	10
B	8	8	16
C	18	12	30
D	6	14	20
E	16	4	20

Profili riga

	I	II	Tot.
A	70	30	100
B	50	50	100
C	60	40	100
D	30	70	100
E	80	20	100



Caso a tre dimensioni



Scelta della metrica

METRICA EUCLIDEA

$$d(i,i') = \sqrt{\sum_{j=1}^{c}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2}$$

$$d(A,B) = \sqrt{(0.585-0.646)^2 + (0.314-0.197)^2 + (0.059-0.061)^2 + (0.042-0.096)^2}$$
$$= 0.143$$

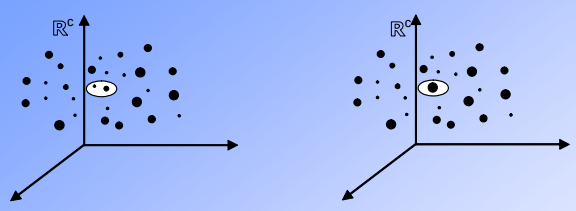
Eccessiva importanza alle modalità con un forte campo di variazione e minor peso a quelle per le quali variazioni sono minori

METRICA DEL CHI-QUADRATO

$$d(i,i') = \sqrt{\sum_{j=1}^{c}\frac{1}{f_{.j}}\left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2}$$

$$d(A,B) = \sqrt{\frac{(0.585-0.646)^2}{0.642} + \frac{(0.314-0.197)^2}{0.203} + \frac{(0.059-0.061)^2}{0.045} + \frac{(0.042-0.096)^2}{0.110}}$$
$$= 0.316$$

Proprietà metrica chi-quadrato



EQUIVALENZA DISTRIBUTIVA

- Il raggruppamento di due profili riga uguali o proporzionali in un unico profilo con massa pari alla somma delle masse, non altera la configurazione dei punti in R^c né la struttura delle distanze in R^r
- Garantisce l'invariabilità dei risultati indipendentemente da come le variabili siano state originariamente codificate
  - non si ha alcuna perdita di informazione se si aggregano modalità
  - non si ha alcuna guadagno di informazione se si suddividono categorie omogenee

$$d(i,i') = \sqrt{\sum_{j=1}^{c} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2}$$

ATTENZIONE ALLE MODALITA' CON FREQUENZE MOLTO BASSE

AC - Calcolo dei fattori

Un qualsiasi studio multidimensionale può essere definito a partire dalla *tripletta* (**X**, **M**, **D**)

AC

matrice dei dati

matrice dei profili (riga o colonna)

metrica

masse dei punti

sistema di pesi per le unità

metrica del *chi-quadrato*

Proiezioni degli r punti sull'asse u $\hat{\mathbf{c}} = \mathbf{D}_r^{-1}\mathbf{F}\,\mathbf{D}_c^{-1}\mathbf{u}$

Obiettivo
$$\begin{cases} \hat{\mathbf{c}}'\,\mathbf{D}_r\,\hat{\mathbf{c}} = \mathbf{u}'\,\mathbf{D}_c^{-1}\mathbf{F}'\,\mathbf{D}_r^{-1}\mathbf{F}\,\mathbf{D}_c^{-1}\mathbf{u} = max \\ \mathbf{u}'\,\mathbf{D}_c^{-1}\mathbf{u} = 1 \end{cases}$$

Soluzione
$$\mathbf{F}'\,\mathbf{D}_r^{-1}\mathbf{F}\,\mathbf{D}_c^{-1}\mathbf{u} = \lambda\mathbf{u}$$

Soluzioni AC e relazioni di transizione

in $\mathbf{R}^r$		in $\mathbf{R}^c$
$\mathbf{S} = \mathbf{F}'\mathbf{D}_r^{-1}\mathbf{F}\mathbf{D}_c^{-1}$	Matrice da diagonalizzare	$\mathbf{S}^* = \mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_r^{-1}$
$\mathbf{u}_\alpha$	Autovettori	$\mathbf{v}_\alpha$
$\varphi_\alpha = \mathbf{D}_c^{-1}\mathbf{u}_\alpha$	Fattori	$\psi_\alpha = \mathbf{D}_r^{-1}\mathbf{v}_\alpha$
$\hat{\psi}_\alpha = \mathbf{D}_r^{-1}\mathbf{F}\varphi_\alpha$	Coordinate	$\hat{\varphi}_\alpha = \mathbf{D}_c^{-1}\mathbf{F}'\psi_\alpha$

$$\hat{\psi}_\alpha = \mathbf{D}_r^{-1}\mathbf{F}\varphi_\alpha = \sqrt{\lambda_\alpha}\ \psi_\alpha$$

$$\hat{\psi}_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{c} \frac{f_{ij}}{f_{i.}} \cdot \hat{\varphi}_{\alpha j}$$

$$\hat{\varphi}_\alpha = \mathbf{D}_c^{-1}\mathbf{F}'\psi_\alpha = \sqrt{\lambda_\alpha}\ \varphi_\alpha$$

$$\hat{\varphi}_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{r} \frac{f_{ij}}{f_{.j}} \cdot \hat{\psi}_{i\alpha}$$

Coordinate delle righe

	(1)	(2)	(3)
A	-0,33	-0,02	0,02
B	-0,04	0,05	-0,06
C	-0,06	-0,05	0,02
D	-0,04	-0,03	-0,01
E	0,10	0,10	0,05
F	0,21	0,02	0,01
G	0,40	-0,15	-0,01

Num.	Autovalore	Perc.	Perc. cum.	Istogramma
1	0.0260	84.23	84.23	••••••••••••••••••••
2	0.0038	12.36	96.59	•••
3	0.0011	3.41	100.00	•

Coordinate delle colonne

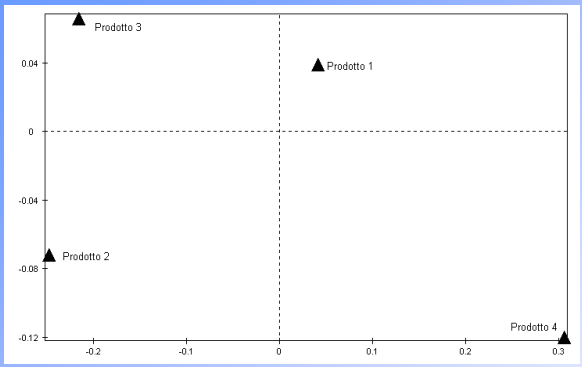
	(1)	(2)	(3)
P1	0,04	0,04	0,01
P2	-0,25	-0,07	0,01
P3	-0,22	0,07	-0,14
P4	0,31	-0,12	-0,03

Ricostruzione della matrice

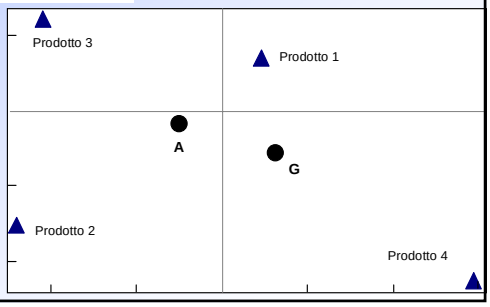
$$f_{ij} = f_{i.}\ f_{.j} \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\ \varphi_{\alpha j}\ \psi_{\alpha i}$$

$$f_{ij} = f_{i.}\ f_{.j} \left(1 + \sum_{\alpha=2}^{p} \sqrt{\lambda_\alpha}\ \varphi_{\alpha j}\ \psi_{\alpha i}\right)$$

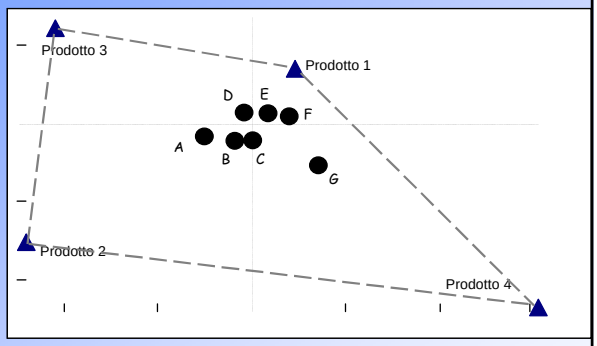
Rappresentazioni

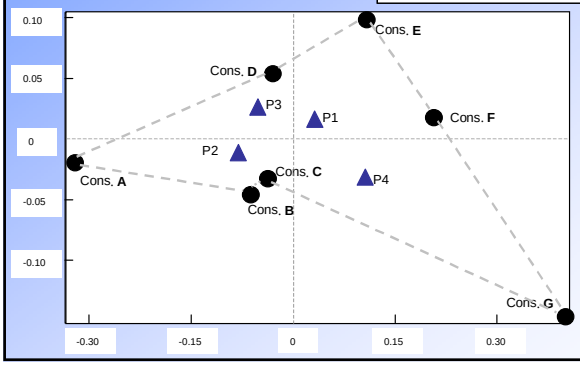


	P1	P2	P3	P4	Totale
A	58.5	31.4	5.9	4.2	*100.0*
B	64.6	19.7	6.1	9.6	*100.0*
C	61.6	23.6	4.3	10.5	*100.0*
D	62.1	22.3	4.7	10.9	*100.0*
E	70.9	15.1	3.5	10.5	*100.0*
F	68.4	13.5	3.2	14.8	*100.0*
G	64.5	11.3	1.6	22.6	*100.0*



Rappresentazioni baricentriche





Rappresentazione congiunta

$$\hat{\varphi} = \mathbf{D}_c^{-1}\mathbf{F}'\hat{\psi} \qquad \text{con} \qquad \hat{\varphi}_j = \sum_i \frac{f_{ij}}{f_{.j}} \cdot \hat{\psi}_i$$

$$\hat{\psi} = \mathbf{D}_r^{-1}\mathbf{F}\hat{\varphi} \qquad \text{con} \qquad \hat{\psi}_i = \sum_j \frac{f_{ij}}{f_{i.}} \cdot \hat{\varphi}_j$$

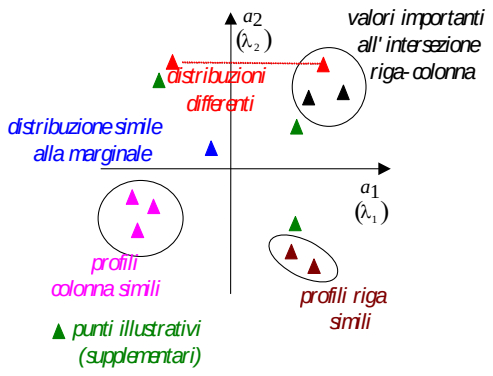
$$\begin{cases} \hat{\varphi} = \beta \cdot \mathbf{D}_c^{-1}\mathbf{F}'\,\hat{\psi} \\ \hat{\psi} = \beta \cdot \mathbf{D}_r^{-1}\mathbf{F}\hat{\varphi} \end{cases}$$

$$\Rightarrow \quad \beta = \frac{1}{\sqrt{\lambda}}$$

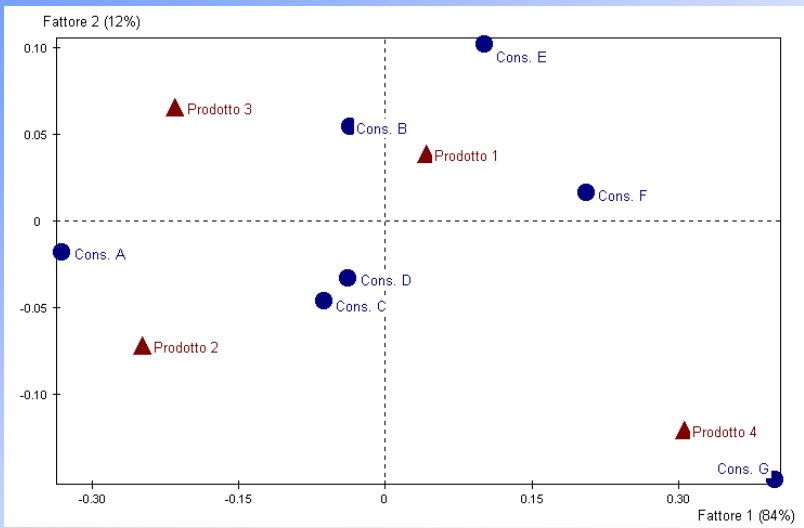
Rappresentazione Simultanea

Analisi delle Corrispondenze

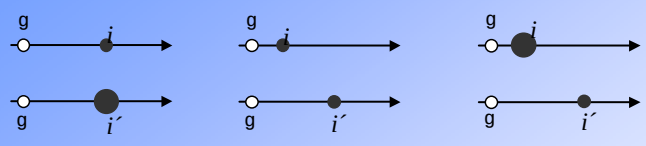
→ lettura prossimità



Rappresentazione congiunta



•Ausili all'interpretazione



$f_{i.} < f_{i'.}$	$f_{i.} = f_{i'.}$	$f_{i.} > f_{i'.}$
$\hat{\psi}_i = \hat{\psi}_{i'}$	$\hat{\psi}_i < \hat{\psi}_{i'}$	$\hat{\psi}_i < \hat{\psi}_{i'}$
$Ca(i) < Ca(i')$	$Ca(i) < Ca(i')$	$Ca(i) = Ca(i')$

$$d^2(i,G) = \sum_{j=1}^{c} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 \quad ; \quad d^2(j,H) = \sum_{i=1}^{r} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2$$

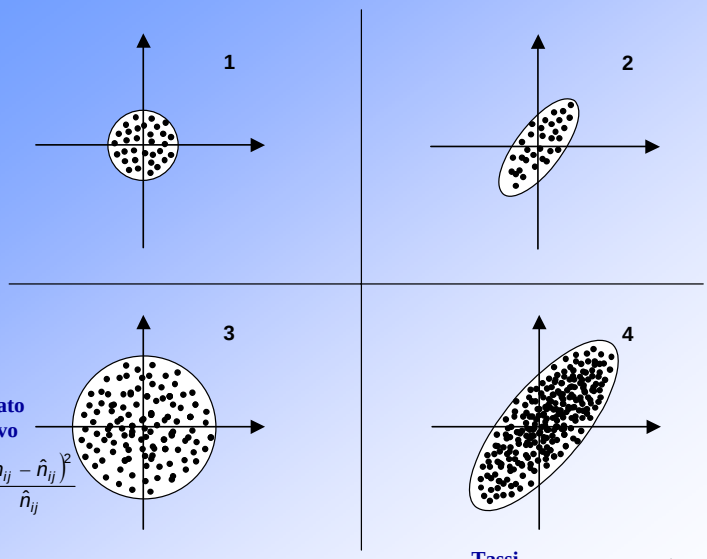
$$Cr_\alpha(i) = \frac{\hat{\psi}_{i\alpha}^2}{d^2(i,G)} \quad ; \quad Cr_\alpha(j) = \frac{\hat{\varphi}_{\alpha j}^2}{d^2(i,H)}$$

$$\sum_\alpha Cr_\alpha(i) = \sum_\alpha Cr_\alpha(j) = 1$$

•Ausili all'interpretazione

	Contributi assoluti			Coseni al quadrato		
	(1)	(2)	(3)	(1)	(2)	(3)
Cat. **A**	0.393	0.008	0.053	0.99	0.00	0.01
Cat. **B**	0.009	0.141	0.605	0.16	0.38	0.45
Cat. **B**	0.033	0.123	0.051	0.62	0.34	0.04
Cat. **D**	0.011	0.058	0.008	0.56	0.42	0.02
Cat. **E**	0.054	0.373	0.273	0.45	0.46	0.09
Cat. **F**	0.200	0.009	0.004	0.99	0.01	0.00
Cat. **G**	0.300	0.288	0.007	0.88	0.12	0.00
P1	0.043	0.254	0.062	0.52	0.45	0.03
P2	0.482	0.278	0.036	0.92	0.08	0.00
P3	0.081	0.051	0.823	0.66	0.06	0.27
P4	0.394	0.417	0.079	0.86	0.13	0.01

Significatività Chi- quadrato e autovalori relativi (*tassi*)



Chi-quadrato significativo

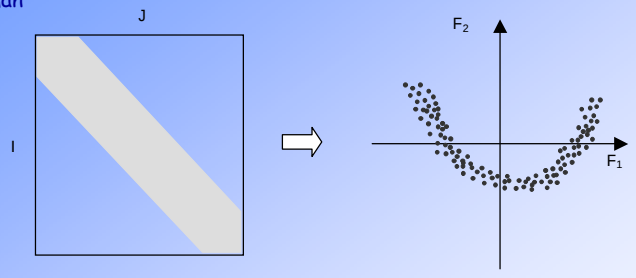
$$X = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Tassi significativi

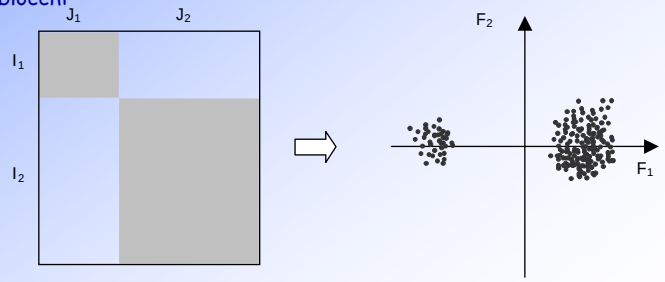
$$\tau_1 = \lambda_1 / \lambda_1 + .. + \lambda_p$$

Alcuni casi particolari

Effetto Guttman



Caso matrici a blocchi



esempio: le preferenze degli spettacoli

	Prosa	Teatro dialettale	Lirica e balletti	Musica classica	Operetta	Rivista e Comm.	Musica leggera	Burattini e marion.	Saggi culturali	Totale
Piemonte	4071	419	583	1121	59	170	1597	250	176	**8446**
Valle d'Aosta	98	12	6	51	2	3	51	0	0	**223**
Lombardia	9598	920	809	2731	96	509	3968	624	318	**19573**
Trentino-A.A.	2171	131	75	651	25	39	857	16	53	**4018**
Veneto	4090	157	505	1546	40	97	1368	115	93	**8011**
Friuli-V.G.	1295	125	138	452	29	18	330	75	53	**2515**
Liguria	1568	197	199	604	32	104	549	60	52	**3365**
Emilia Rom.	5798	1052	562	1669	115	198	2009	239	144	**11786**
Toscana	4315	85	574	1709	48	162	1310	283	171	**8657**
Umbria	1001	59	132	403	13	17	147	28	36	**1836**
Marche	1991	190	155	579	20	74	480	85	66	**3640**
Lazio	12940	281	1006	2051	40	1023	1040	382	208	**18971**
Abruzzo	1033	22	72	705	22	8	283	27	25	**2197**
Molise	102	10	10	60	1	3	24	3	2	**215**
Campania	3346	1041	283	589	23	120	749	138	187	**6476**
Puglia	2013	311	211	896	21	72	504	69	49	**4146**
Basilicata	293	29	23	114	3	11	91	12	0	**576**
Calabria	721	188	73	448	1	10	199	15	75	**1730**
Sicilia	3770	351	300	1162	42	171	658	156	159	**6769**
Sardegna	1339	116	109	754	18	44	398	157	89	**3024**
Totale	**61553**	**5696**	**5825**	**18295**	**650**	**2853**	**16612**	**2734**	**1956**	**116174**

Test di indipendenza:

VALORE DEL CHI-QUADRATO	=	10102.542
GRADI DI LIBERTA'	=	152.000
P-VALUE	=	.000

esempio: i profili riga

	Prosa	Teatro dialettale	Lirica e balletti	Musica classica	Operetta	Rivista e Comm.	Musica leggera	Burattini e marion.	Saggi culturali	Totale
Piemonte	48.2	5.0	6.9	13.3	0.7	2.0	18.9	3.0	2.1	100.0
Valle d'Aosta	43.9	5.4	2.7	22.9	0.9	1.3	22.9	0.0	0.0	100.0
Lombardia	49.0	4.7	4.1	14.0	0.5	2.6	20.3	3.2	1.6	100.0
Trentino-A.A.	54.0	3.3	1.9	16.2	0.6	1.0	21.3	0.4	1.3	100.0
Veneto	51.1	2.0	6.3	19.3	0.5	1.2	17.1	1.4	1.2	100.0
Friuli-V.G.	51.5	5.0	5.5	18.0	1.2	0.7	13.1	3.0	2.1	100.0
Liguria	46.6	5.9	5.9	17.9	1.0	3.1	16.3	1.8	1.5	100.0
Emilia Rom.	49.2	8.9	4.8	14.2	1.0	1.7	17.0	2.0	1.2	100.0
Toscana	49.8	1.0	6.6	19.7	0.6	1.9	15.1	3.3	2.0	100.0
Umbria	54.5	3.2	7.2	21.9	0.7	0.9	8.0	1.5	2.0	100.0
Marche	54.7	5.2	4.3	15.9	0.5	2.0	13.2	2.3	1.8	100.0
Lazio	68.2	1.5	5.3	10.8	0.2	5.4	5.5	2.0	1.1	100.0
Abruzzo	47.0	1.0	3.3	32.1	1.0	0.4	12.9	1.2	1.1	100.0
Molise	47.4	4.7	4.7	27.9	0.5	1.4	11.2	1.4	0.9	100.0
Campania	51.7	16.1	4.4	9.1	0.4	1.9	11.6	2.1	2.9	100.0
Puglia	48.6	7.5	5.1	21.6	0.5	1.7	12.2	1.7	1.2	100.0
Basilicata	50.9	5.0	4.0	19.8	0.5	1.9	15.8	2.1	0.0	100.0
Calabria	41.7	10.9	4.2	25.9	0.1	0.6	11.5	0.9	4.3	100.0
Sicilia	55.7	5.2	4.4	17.2	0.6	2.5	9.7	2.3	2.3	100.0
Sardegna	44.3	3.8	3.6	24.9	0.6	1.5	13.2	5.2	2.9	100.0
Profilo medio	***53.0***	***4.9***	***5.0***	***15.7***	***0.6***	***2.5***	***14.3***	***2.4***	***1.7***	***100.0***

esempio: i profili colonna

	Prosa	Teatro dialettale	Lirica e balletti	Musica classica	Operetta	Rivista e Comm.	Musica leggera	Burattini e marion.	Saggi culturali	*Profilo medio*
Piemonte	6.6	7.4	10.0	6.1	9.1	6.0	9.6	9.1	9.0	*7.3*
Valle d'Aosta	0.2	0.2	0.1	0.3	0.3	0.1	0.3	0.0	0.0	*0.2*
Lombardia	15.6	16.2	13.9	14.9	14.8	17.8	23.9	22.8	16.3	*16.8*
Trentino-A.A.	3.5	2.3	1.3	3.6	3.8	1.4	5.2	0.6	2.7	*3.5*
Veneto	6.6	2.8	8.7	8.5	6.2	3.4	8.2	4.2	4.8	*6.9*
Friuli-V.G.	2.1	2.2	2.4	2.5	4.5	0.6	2.0	2.7	2.7	*2.2*
Liguria	2.5	3.5	3.4	3.3	4.9	3.6	3.3	2.2	2.7	*2.9*
Emilia Romagna	9.4	18.5	9.6	9.1	17.7	6.9	12.1	8.7	7.4	*10.1*
Toscana	7.0	1.5	9.9	9.3	7.4	5.7	7.9	10.4	8.7	*7.5*
Umbria	1.6	1.0	2.3	2.2	2.0	0.6	0.9	1.0	1.8	*1.6*
Marche	3.2	3.3	2.7	3.2	3.1	2.6	2.9	3.1	3.4	*3.1*
Lazio	21.0	4.9	17.3	11.2	6.2	35.9	6.3	14.0	10.6	*16.3*
Abruzzo	1.7	0.4	1.2	3.9	3.4	0.3	1.7	1.0	1.3	*1.9*
Molise	0.2	0.2	0.2	0.3	0.2	0.1	0.1	0.1	0.1	*0.2*
Campania	5.4	18.3	4.9	3.2	3.5	4.2	4.5	5.0	9.6	*5.6*
Puglia	3.3	5.5	3.6	4.9	3.2	2.5	3.0	2.5	2.5	*3.6*
Basilicata	0.5	0.5	0.4	0.6	0.5	0.4	0.5	0.4	0.0	*0.5*
Calabria	1.2	3.3	1.3	2.4	0.2	0.4	1.2	0.5	3.8	*1.5*
Sicilia	6.1	6.2	5.2	6.4	6.5	6.0	4.0	5.7	8.1	*5.8*
Sardegna	2.2	2.0	1.9	4.1	2.8	1.5	2.4	5.7	4.6	*2.6*
Totale	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*

esempio: la diagonalizzazione

Num.	Autovalore	Perc.	Perc. cum.	Istogramma
1	0.03904	44.90	44.90	••••••••••••••••••••••••
2	0.02644	30.41	75.30	••••••••••••••••
3	0.01224	14.08	89.38	•••••••
4	0.00417	4.80	94.18	•••
5	0.00253	2.92	97.10	••
6	0.00127	1.47	98.56	•
7	0.00071	1.03	99.59	•
8	0.00035	0.41	100.00	

Numero degli assi (rappresentazione ottimale)

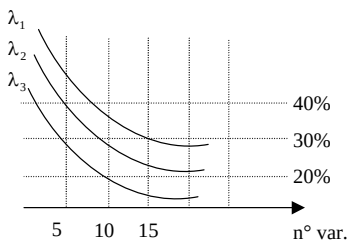
- percentuale varianza $\tau_h = \lambda_1 + ... + \lambda_h / \lambda_1 + ... + \lambda_p \geq 0.75$

- scree test:

  (max caduta istogr. autovalori)

λ_1 ★ ★ ★★★★★★★★
λ_2 ★ ★ ★★★★★★
λ_3 ★ ★

- abachi



ACB

num. autovalori tale che:

$\tau_h > \tau_{tabulati}$

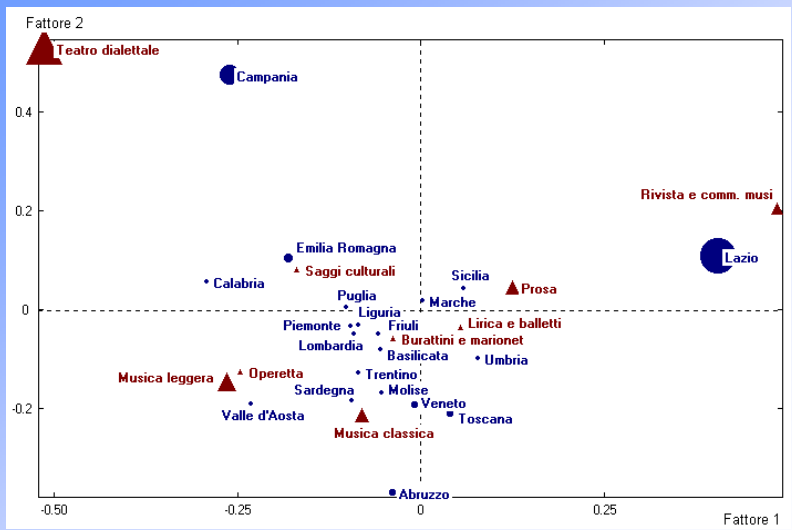
esempio: contributi spettacoli

	Contributi assoluti (x100)						Coseni al quadrato					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Prosa	21,10	4,30	0,10	6,20	0,20	5,80	0,84	0,12	0,00	0,03	0,00	0,01
Teatro dialettale	33,30	53,00	4,20	0,20	0,00	3,40	0,47	0,51	0,02	0,00	0,00	0,00
Lirica e balletti	0,40	0,20	0,00	11,00	79,20	0,30	0,06	0,02	0,00	0,17	0,74	0,00
Musica classica	2,60	26,40	43,20	0,30	2,40	5,60	0,08	0,52	0,39	0,00	0,01	0,01
Operetta	0,90	0,30	0,00	0,40	2,70	9,80	0,30	0,08	0,00	0,02	0,06	0,11
Rivista e Comm.	14,90	4,00	4,30	2,10	5,20	23,50	0,71	0,13	0,06	0,01	0,02	0,04
Musica leggera	25,50	11,10	43,80	2,10	0,20	0,30	0,54	0,16	0,29	0,01	0,00	0,00
Bur. e marion.	0,10	0,30	1,80	58,70	7,70	5,70	0,01	0,02	0,07	0,76	0,06	0,02
Saggi culturali	1,20	0,40	2,50	19,20	2,30	45,60	0,20	0,05	0,13	0,33	0,02	0,24

esempio: contributi regioni

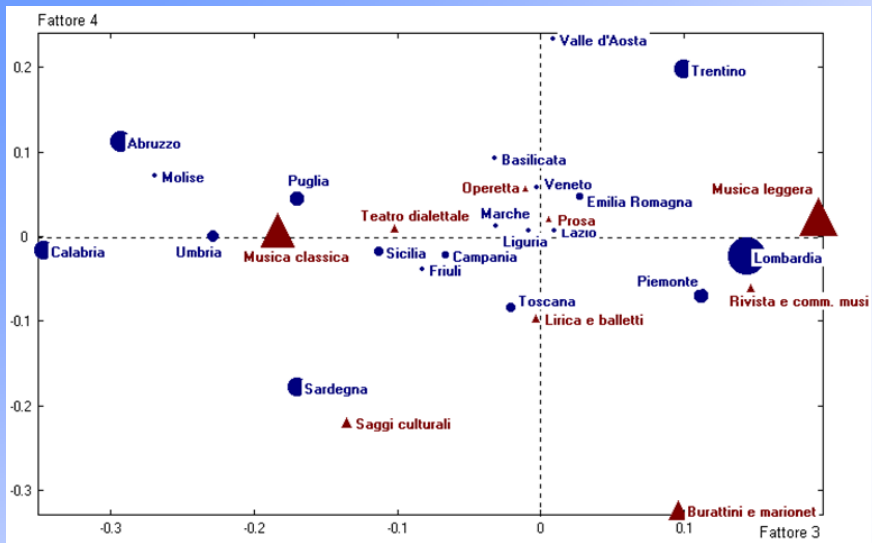
	Contributi assoluti (x100)						Coseni al quadrato					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Piemonte	1,70	0,30	7,50	8,20	16,60	2,30	0,27	0,03	0,37	0,14	0,17	0,01
Valle d'Aosta	0,30	0,30	0,00	2,50	0,10	1,20	0,34	0,23	0,00	0,35	0,01	0,05
Lombardia	3,60	1,40	28,70	2,00	17,20	0,80	0,24	0,06	0,60	0,01	0,08	0,00
Trentino-A.A.	0,60	2,10	2,80	32,70	7,50	20,20	0,09	0,19	0,12	0,46	0,06	0,09
Veneto	0,00	9,40	0,00	5,70	14,70	3,70	0,00	0,77	0,00	0,07	0,12	0,02
Friuli-V.G.	0,20	0,20	1,20	0,70	1,00	0,70	0,13	0,09	0,26	0,05	0,05	0,02
Liguria	0,50	0,10	0,00	0,00	1,90	13,10	0,32	0,04	0,00	0,00	0,07	0,25
Emilia Rom.	8,40	4,20	0,60	5,80	3,20	15,90	0,34	0,22	0,02	0,05	0,02	0,04
Toscana	0,30	12,40	3,00	12,40	4,00	1,40	0,03	0,80	0,01	0,13	0,03	0,00
Umbria	0,20	0,60	6,70	0,00	7,00	1,70	0,08	0,12	0,64	0,00	0,14	0,02
Marche	0,00	0,00	0,30	0,10	0,80	1,30	0,00	0,11	0,27	0,05	0,18	0,14
Lazio	69,10	7,50	0,10	0,30	0,30	0,50	0,93	0,07	0,00	0,00	0,00	0,00
Abruzzo	0,10	9,80	13,30	5,80	2,80	3,30	0,01	0,56	0,35	0,05	0,02	0,01
Molise	0,00	0,20	1,10	0,20	0,00	0,90	0,02	0,23	0,61	0,05	0,00	0,05
Campania	9,60	47,70	2,00	0,60	0,00	4,70	0,23	0,76	0,01	0,00	0,00	0,00
Puglia	1,00	0,00	8,40	1,80	0,30	12,00	0,22	0,00	0,60	0,04	0,00	0,09
Basilicata	0,00	0,10	0,00	1,10	0,10	3,30	0,09	0,19	0,03	0,27	0,02	0,26
Calabria	3,30	0,20	14,70	0,10	1,30	9,50	0,36	0,01	0,51	0,00	0,01	0,03
Sicilia	0,50	0,40	6,00	0,40	2,20	2,60	0,17	0,09	0,61	0,01	0,05	0,03
Sardegna	0,60	3,30	6,20	19,60	18,90	1,00	0,07	0,27	0,24	0,26	0,15	0,00

esempio: rappresentazione congiunta



Dimensioni punti proporzionali al contributo assoluto

esempio: rappresentazione congiunta



Analisi delle Corrispondenze Multiple (ACM)

- E' uno dei più importanti metodi per l'analisi di variabili qualitative o miste;
- Ha un vastissimo campo di applicazione e si rivela particolarmente adatta all'analisi di dati di inchiesta ed alla descrizione di tabelle di grandi dimensioni;
- Le righe della tabella sono in generale individui o osservazioni;
- Le colonne della tabella sono le modalità delle variabili nominali (o le classi di valori delle variabili continue) e rappresentano spesso le modalità di risposta relative alle domande di un questionario.

• I dati e la codifica

p

R = (*i*)

sex (2)	età (3)						prof (4)
2	1	4	4	3	1	3	1

Questionario

n

Codifica ridotta

p variabili
s modalità

$s = s_1+s_2+\ldots+s_p$

Codifica disgiuntiva completa

Z =

m	f	<30	30-50	>50		stu	imp	dir	pen
0	1	1	0	0		1	0	0	0

s_1 s_2 . . . s_p

• L'indagine sui laureati in Economia

V1	Genere	2 mod.	Maschio / Femmina
V2	Residenza	3 mod.	Na città / Na prov. / Altre prov.
V3	Voto di laurea	4 mod.	<96 / 96-105 / 106-110 / 110 e l.
V4	Età attuale	*continua*	
V5	Tipo diploma	4 mod.	Classico / Scientif. / Tecnico / Altro
V6	Voto diploma	4 mod.	36-42 / 43-48 / 49-54 / 55-60
V7	Frequenza	4 mod.	Solo esami / <30% lez. / 30-50% / >50%
V8	Materia tesi	5 mod.	Econom. / Giurid. / Quantit. / Aziend. / Altro
V9	Borse di studio post-laurea	2 mod.	Si / No
V10	Abilitazione alla prof.	2 mod.	Si / No
V11	Lavoro durante gli studi	3 mod.	Lav. Stabile / Part-time / Stud. Tempo pieno
V12	Posizione occupazionale	4 mod.	Non occ. / Non occ. per sc. / Precario / Occ. Stabile

La matrice dei dati (*Codifica ridotta*)

	Id.	V_1	V_2	V_3	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
	1	1	3	2	2	2	2	1	2	1	3	4
	2	2	3	2	2	1	2	5	2	2	3	1
	3	1	1	2	1	2	3	1	2	2	3	4
	4	1	2	2	2	3	3	2	1	2	3	3
R =	**5**	1	3	1	2	1	3	2	2	2	3	1
	:	:	:	:	:	:	:	:	:	:	:	:
	:	:	:	:	:	:	:	:	:	:	:	:
	383	2	3	2	1	2	3	3	1	2	3	3
	384	2	3	3	2	4	2	4	1	1	3	1
	385	2	3	4	2	4	3	4	1	1	3	3

La matrice dei dati (*Codifica disgiuntiva completa*)

		V_1		V_2			V_3				...			V_{11}			V_{12}			
	1	1	0	0	0	1	0	1	0	0				0	0	1	0	0	0	1
	2	0	1	0	0	1	0	1	0	0				0	0	1	1	0	0	0
	3	1	0	1	0	0	0	1	0	0				0	0	1	0	0	0	1
	4	1	0	0	1	0	1	0	0	0				0	0	1	0	0	1	0
Z =	**5**	1	0	0	0	1	1	0	0	0				0	0	1	1	0	0	0
	:																			
	:																			
	383	0	1	0	0	1	0	1	0	0				0	0	1	0	0	1	0
	384	0	1	0	0	1	0	0	1	0				0	0	1	1	0	0	0
	385	0	1	0	0	1	0	0	0	1				0	0	1	0	0	1	0

s=37

• La matrice da diagonalizzare (tabella di Burt)

➡ ***Z'Z***

	Ma	Fe	NaCi	NaPr	AlPr	…	Occ1	Occ2	Occ3	Occ4
Ma	226	0								
Fe	0	159								
NaCi	116	65	181	0	0					
NaPr	65	50	0	115	0					
AlPr	45	44	0	0	89					
:						…				
Occ1	49	60	51	32	26		109	0	0	0
Occ2	10	13	10	6	7		0	23	0	0
Occ3	72	48	48	37	35		0	0	120	0
Occ4	95	38	72	40	21		0	0	0	133

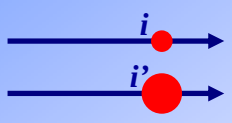
Gli autovalori – nell'ACM si usa lo *scree test*

Num.	Autov.	%	% cum.	Istogramma
1	0.2108	8.92	8.92	••••••••••••••••••••••••••••••••••••••
2	0.1554	6.58	15.50	••••••••••••••••••••••••••••
3	0.1333	5.64	21.14	••••••••••••••••••••••••
4	0.1219	5.16	26.29	••••••••••••••••••••••••
5	0.1156	4.89	31.18	••••••••••••••••••••••••
6	0.1105	4.68	35.86	••••••••••••••••••••••
7	0.1055	4.46	40.32	••••••••••••••••••••
8	0.1026	4.34	44.67	••••••••••••••••••••
9	0.0979	4.14	48.81	••••••••••••••••••
10	0.0968	4.10	52.90	••••••••••••••••••
11	0.0936	3.96	56.86	••••••••••••••••••
12	0.0920	3.89	60.76	••••••••••••••••••
13	0.0862	3.65	64.40	••••••••••••••••
14	0.0841	3.56	67.96	••••••••••••••••
15	0.0796	3.37	71.33	••••••••••••••••
16	0.0774	3.27	74.60	••••••••••••••
17	0.0764	3.23	77.84	••••••••••••••
18	0.0735	3.11	80.94	••••••••••••••
19	0.0702	2.97	83.91	••••••••••••••
20	0.0659	2.79	86.70	••••••••••••
21	0.0624	2.64	89.34	••••••••••••
22	0.0578	2.44	91.79	••••••••••
23	0.0563	2.38	94.17	••••••••••
24	0.0533	2.26	96.42	•••••••••
25	0.0442	1.87	98.29	•••••••
26	0.0403	1.71	100.00	•••••••

• La lettura dei risultati

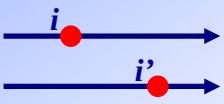
1. Gli autovalori sono una misura eccessivamente pessimistica dell'effettivo potere esplicativo dei fattori individuati;

2 I fattori vanno interpretati tenendo conto dei valori dei contributi assoluti e non delle coordinate dei punti (come avveniva in ACP)



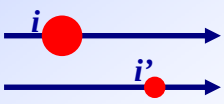
$Cr(i)<Cr(i')$

$f(i.)<f(i'.)$



$Cr(i)<Cr(i')$

$\psi(i)< \psi(i')$



$Cr(i)=Cr(i')$

$f(i.)\psi^2(i)= f(i'.)\psi^2(i')$

• Qualche nota

n = individui; p = variabili; s = modalità
$z_{.j}$ = freq. della j-esima modalità

Numero di autovalori → $s - p$

Tassi di inerzia → $\frac{\lambda_\alpha}{\sum_{j=1}^{s-p} \lambda_\alpha} \times 100$

Distanza di una modalità dal baricentro → $d^2(j,G) = \frac{n}{z_{.j}} - 1$

Inerzia di una modalità → $I(j) = m_j \times d^2(j,G)$ con $m_j = \frac{z_j}{n \times p}$

$I(j) = \frac{1}{p} \cdot \left(1 - \frac{z_{.j}}{n}\right)$

ATTENZIONE alle modalità con frequenze molto basse!!!

Inerzia totale → $I = \frac{s}{p} - 1$

• Esempio

```
ANALYSE DES CORRESPONDANCES MULTIPLES
APUREMENT DES MODALITES ACTIVES
SEUIL (PCMIN)    :       2.00 %        POIDS:        1.62
AVANT APUREMENT :      13 QUESTIONS ACTIVES          37 MODALITES ASSOCIEES
APRES           :      13 QUESTIONS ACTIVES          37 MODALITES ASSOCIEES
POIDS TOTAL DES INDIVIDUS ACTIFS :      81.00
TRI-A-PLAT DES QUESTIONS ACTIVES
----------------------------+-----------------+-------------------------------------------------------------
         MODALITES          | AVANT APUREMENT |            APRES APUREMENT
IDENT      LIBELLE          |  EFF.     POIDS |  EFF.     POIDS       HISTOGRAMME DES POIDS RELATIFS
----------------------------+-----------------+-------------------------------------------------------------
    1 . genere
masc - maschio              |   23      23.00 |   23      23.00  oooooooooooooooooo
femm - femmina              |   58      58.00 |   58      58.00  oooooooooooooooooooooooooooooooooooooooooooooo
----------------------------+-----------------+-------------------------------------------------------------
    2 . tipo diploma
lice - liceo                |   24      24.00 |   24      24.00  oooooooooooooooooo
ITC  - ITC                  |   46      46.00 |   46      46.00  oooooooooooooooooooooooooooooooooo
Aldp - Altro diploma        |   11      11.00 |   11      11.00  ooooooooo
----------------------------+-----------------+-------------------------------------------------------------
    3 . voto diploma
vot1 - 36-42                |   14      14.00 |   14      14.00  oooooooooo
vot2 - 43-48                |   20      20.00 |   20      20.00  ooooooooooooooo
vot3 - 49-54                |   25      25.00 |   25      25.00  oooooooooooooooooooo
vot4 - 55-60                |   22      22.00 |   22      22.00  ooooooooooooooooo
----------------------------+-----------------+-------------------------------------------------------------
    4 . anno corso
an2  - II anno              |   56      56.00 |   56      56.00  oooooooooooooooooooooooooooooooooooooooooooooo
an3  - III anno             |   18      18.00 |   18      18.00  ooooooooooooooo
an4  - IV anno              |    4       4.00 |    4       4.00  ooo
anfc - fuori corso          |    3       3.00 |    3       3.00  ooo
----------------------------+-----------------+-------------------------------------------------------------
```

In questi casi è meglio ricodificare la variabili accorpando le classi contigue

• La tabella dei contributi

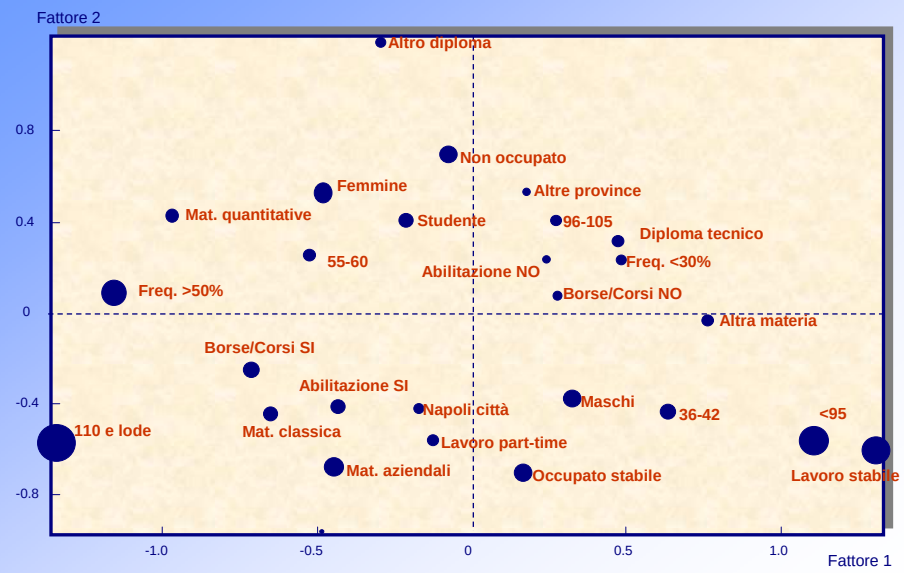
1 . Genere					
Maschi	2.8	4.8	0.0	0.2	2.7
Femmine	4.0	6.8	0.1	0.2	3.9
Contr. Cum.	*6.8*	*11.5*	*0.1*	*0.4*	*6.6*
2 - Residenza					
Napoli città	0.6	4.8	0.4	3.6	1.0
Napoli provincia	0.2	1.1	10.0	2.3	1.2
Altre province	0.3	3.8	7.2	1.0	0.0
Contr. Cum.	*1.1*	*9.7*	*17.7*	*6.9*	*2.2*
3 - Voto di laurea					
<95	9.6	3.4	0.0	1.0	1.5
96-105	1.4	4.4	0.8	0.0	4.2
106-110	1.4	0.2	0.2	5.7	1.9
110 e lode	14.1	3.4	1.2	1.9	0.4
Contr. Cum.	*26.5*	*11.4*	*2.2*	*8.6*	*8.0*
5 – Tipo diploma					
Classico	3.8	2.5	0.0	0.4	1.6
Scientifico	0.1	1.0	6.9	0.0	5.1
Tecnico	3.6	2.1	3.1	0.7	5.7
Altro diploma	0.2	4.1	3.8	0.5	8.9
Contr. Cum.	*7.6*	*9.7*	*13.9*	*1.5*	*21.4*
6 - Voto diploma					
36-42	3.6	2.3	7.1	5.0	6.4
43-48	0.8	0.5	0.0	1.5	12.7
49-54	0.7	1.3	0.0	1.3	0.1
55-60	3.2	1.0	6.5	0.1	2.8
Contr. Cum.	*8.3*	*5.1*	*13.6*	*8.0*	*22.0*
7 - Frequenza ai corsi					
Solo per esami	2.1	0.2	3.0	0.1	4.8
Meno del 30%	3.8	1.2	0.1	0.0	0.0
Tra 30% e 50%	0.4	1.2	3.7	1.9	0.6
Oltre il 50%	8.5	0.1	2.6	4.7	0.1
Contr. Cum.	*14.9*	*2.7*	*9.5*	*6.7*	*5.5*

8 - Argomento tesi					
Mat. Econ.	0.1	1.1	0.1	18.6	1.2
Mat. Giur.	1.4	1.4	0.2	0.4	0.1
Mat. Quant.	3.3	0.9	1.8	13.6	0.5
Mat. Aziend.	2.2	7.0	1.3	0.4	0.0
Altra materia	4.0	0.0	0.5	5.7	1.2
Contr. Cum.	*10.9*	*10.5*	*3.9*	*38.8*	*3.1*
9 - Borse post-laurea					
Borse/corsi si	6.1	0.9	0.0	1.7	0.0
Borse/corsi no	2.4	0.4	0.0	0.7	0.0
Contr. Cum.	*8.5*	*1.3*	*0.0*	*2.3*	*0.0*
10 - Abilitazione					
Abilitazione si	2.9	3.6	0.7	3.0	3.4
Abilitazione no	1.7	2.1	0.4	1.7	1.9
Contr. Cum.	*4.6*	*5.7*	*1.2*	*4.6*	*5.3*
11 - Durante gli studi					
Lavoro stabile	8.9	2.6	17.0	0.0	1.9
Lavoro part-time	0.2	5.4	6.6	0.4	0.8
Studente	1.1	5.8	0.0	0.3	1.7
Contr. Cum.	*10.2*	*13.8*	*23.6*	*0.7*	*4.4*
12 – Condiz. attuale					
Non occupato	0.1	8.1	2.3	5.6	1.4
Non occ. per sc.	0.0	0.1	0.8	6.9	9.2
Precario/Tiroc.	0.2	0.3	1.9	6.9	10.3
Occupato stabile	0.4	10.3	9.4	2.1	0.5
Contr. Cum.	*0.7*	*18.7*	*14.4*	*21.4*	*21.5*

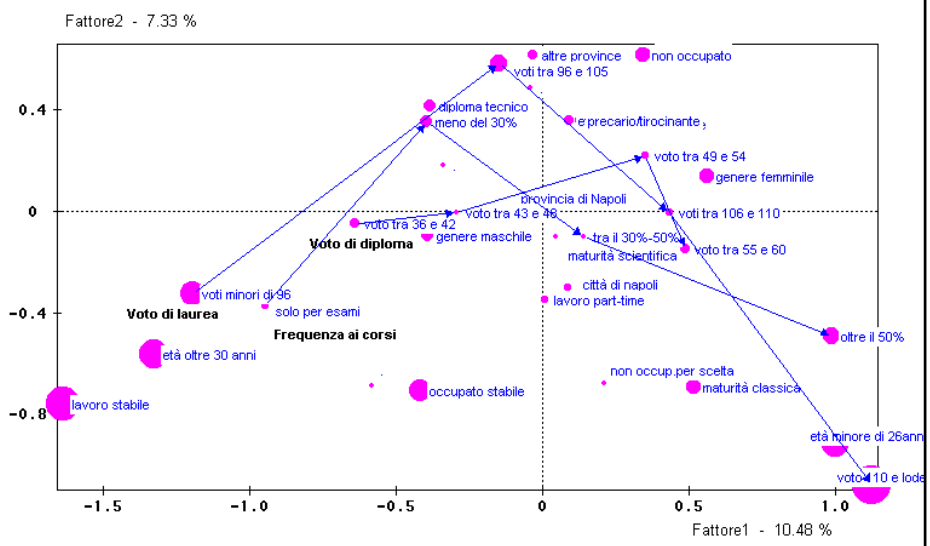
• La lettura dei piani fattoriali

- Ogni ***domanda del questionario*** (variabile) è rappresentata da tanti punti quante sono le sue modalità di risposta
- Il ***baricentro*** della nube dei punti (origine degli assi) è anche il baricentro delle modalità di ciascuna variabile
- La ***prossimità tra due modalità*** esprime una stretta ***connessione*** (interdipendenza) dei temi che esse rappresentano
- Le prossimità tra modalità relative a variabili ***attive*** e ***supplementari*** vanno lette in termini di dipendenza delle seconde rispetto alle prime
- Se le modalità di una variabile sono ordinate, si studieranno le ***traiettorie*** seguite da questi punti, le loro forme e le loro posizioni reciproche
- La ***prossimità tra punti-unità*** (intervistati) evidenzia una posizione simile rispetto ai temi trattati nel questionario
- La ***prossimità tra punti-modalità e punti-unità***, qualora si ricorra a rappresentazioni congiunte, rivelano l'importanza che talune caratteristiche hanno nel profilo di ciascun intervistato o di gruppi di questi

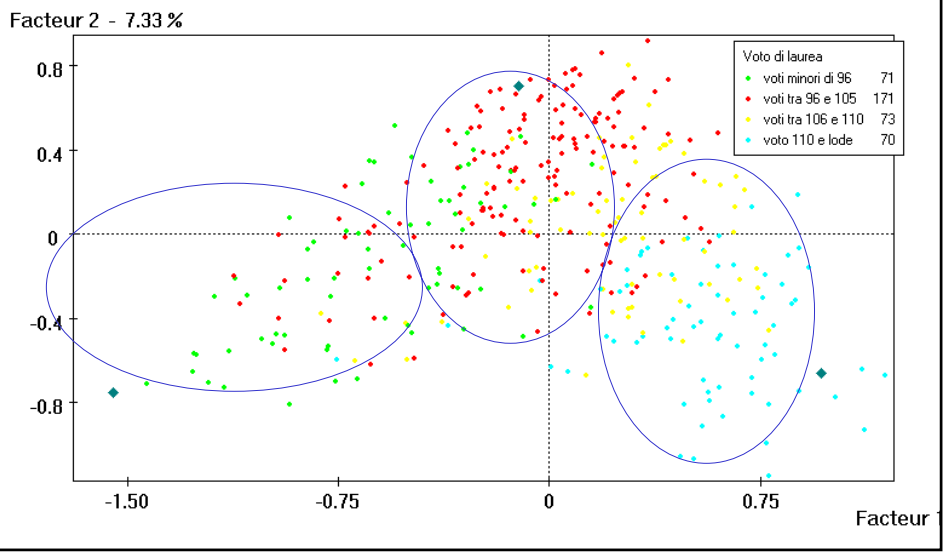
I Laureati in Economia della Federico II
la rappresentazione grafica



Ausilii all'interpretazione: l'inversione degli assi e le traiettorie



Ausilii all'interpretazione : la densità degli individui e il voto



• La descrizione dei fattori con i valori-test

n_j → Numero di soggetti che presentano la modalità **j**

$\hat{\varphi}_{\alpha j}$ → Coordinata della modalità **j** nel caso di estrazione casuale degli n_j soggetti

$$E(\hat{\varphi}_{\alpha j}) = 0$$

$$Var(\hat{\varphi}_{\alpha j}) = \frac{1}{n_j} \cdot \frac{n - n_j}{n - 1}$$

Per n_j sufficientemente grande →

$$\hat{\varphi}_{\alpha j} \cdot \sqrt{n_j \cdot \frac{n-1}{n-n_j}} \sim N(0,1)$$