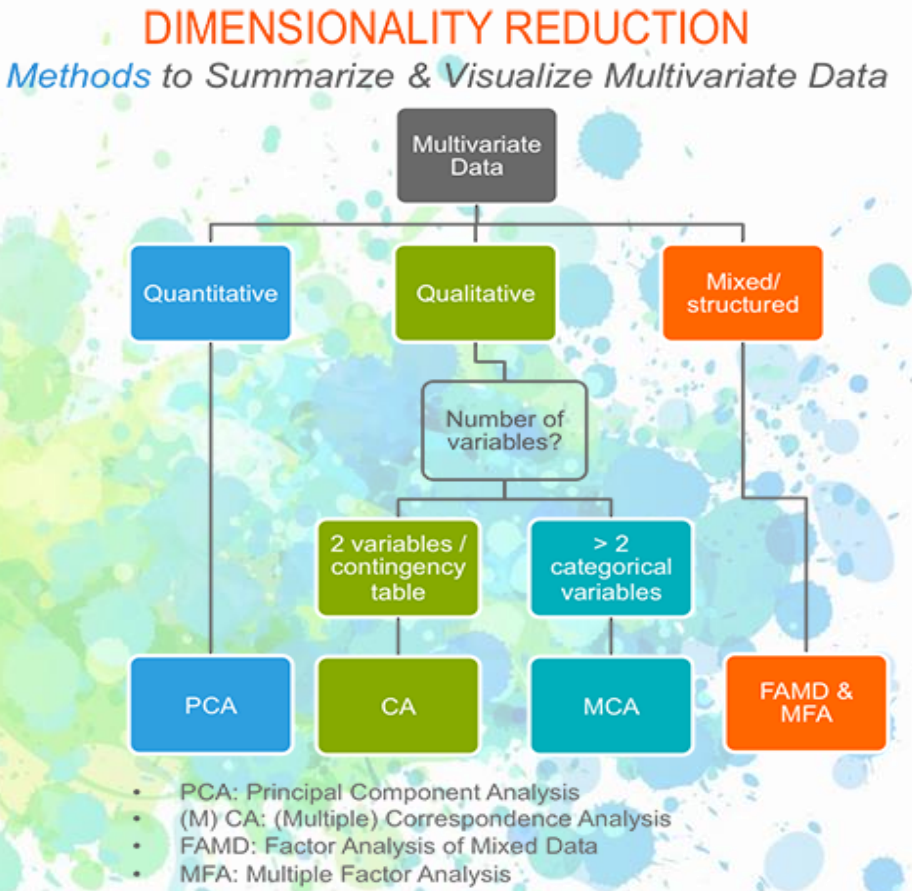
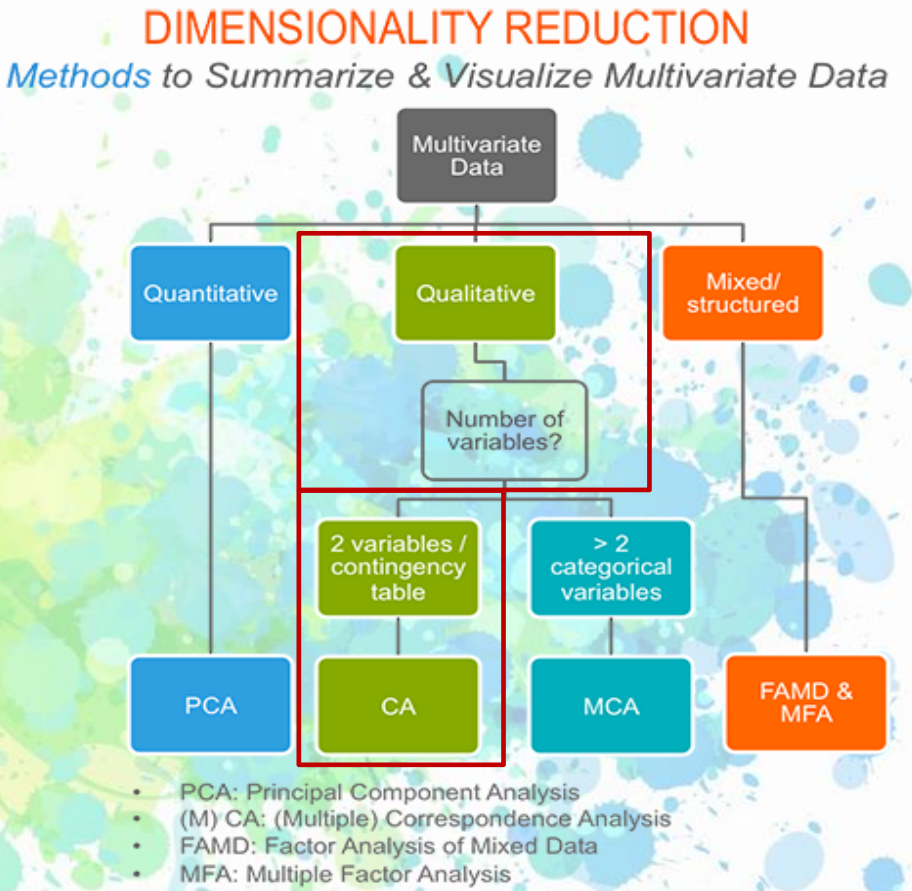
Prof.ssa Rosanna Cataldo



L' ANALISI DELLE CORRISPONDENZE (AC)



L'ANALISI DELLE CORRISPONDENZE (AC)

L'Analisi multivariata di tabelle caratterizzate da variabili di tipo categorico ha avuto una evoluzione più lenta di quanto non sia avvenuto parallelamente nel campo delle variabili numeriche.

Per molti anni si è tentato di utilizzare tecniche nate per variabili numeriche per analizzare variabili categoriche

È solo verso la fine del 1960, in un periodo in cui la metodologia statistica evidenziava un forte ritardo nei confronti delle problematiche derivanti dall'osservazione di fenomeni relativi al campo delle scienze sociali e comportamentali, che la ricerca si indirizza verso l'individuazione di tecniche specifiche per l'analisi simultanea di variabili categoriche

In questi anni **Benzecrì** propone un diverso approccio al problema dello studio dell'associazione tra variabili categoriche. L'approccio viene battezzato dallo stesso Benzecrì ***Analisi delle corrispondenze*** e costituisce il punto di partenza di quello che sarebbe poi divenuto il filone francese di *Analyse des donneèes*

L'ANALISI DELLE CORRISPONDENZE (AC)

È un'estensione dell'ACP adatta per esplorare le relazioni tra variabili qualitative (o dati categoriali). Come l'analisi delle componenti principali, l'AC fornisce una soluzione per riassumere e visualizzare i set di dati in grafici a due dimensioni.

- Se l'insieme include solo due variabili, il metodo è usualmente chiamato **Analisi delle Corrispondenze Semplici (SCA)**
- Se l'analisi coinvolge più di due variabili, allora è usualmente chiamata **Analisi delle Corrispondenze Multiple (MCA)**

Obiettivo → **L'obiettivo è individuare uno spazio di poche dimensioni all'interno del quale collocare le modalità della variabili inserite nell'analisi**

L'Analisi delle Corrispondenze è una tecnica che considera tabelle (generalmente **tabelle di contingenza**) ottenute incrociando due caratteristiche di un fenomeno (generalmente due **variabili qualitative**), caratteristiche delle quali si ha interesse a studiare la distribuzione congiunta.

In particolare, risulta utile quando siamo interessati a studiare le **similarità** fra gli elementi appartenenti a ciascuno dei due insiemi (righe e colonne della tabella), attraverso la rappresentazione fattoriale della configurazione, o forma, delle nubi di punti associate a tali insiemi, rappresentazione che fornisce, dunque, una visione sintetica e globale delle relazioni fra i punti ma anche una lettura analitica dei particolari aspetti di queste relazioni.

Analisi delle Corrispondenze Semplici (SCA) →

LA MATRICE DEI DATI E L'OBIETTIVO DEL METODO

I dati di partenza sono generalmente costituiti da n individui sui quali vengono osservate due variabili qualitative.

ID	Genere	Diploma
1	F	Liceo Classico
2	F	Liceo Scientifico
3	F	Liceo Classico
4	F	Liceo Classico
5	M	Istituto Professionale
6	F	Istituto Professionale
7	M	Liceo Classico
8	M	Liceo Scientifico
9	M	Istituto Professionale
10	M	Istituto Tecnico
11	M	Istituto Tecnico
12	F	Istituto Tecnico
13	F	Liceo Scientifico
14	M	Istituto Tecnico
15	M	Istituto Tecnico
16	M	Liceo Scientifico
17	M	Liceo Scientifico
18	F	Liceo Scientifico
19	F	Liceo Scientifico
20	F	Liceo Scientifico

I dati sono quindi inseriti in **una tabella di contingenza**, in cui le **righe** definiscono le modalità della variabile **Genere** e le **colonne** quelle della variabile **Diploma**

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	1	1	3	5	**10**
M	2	4	1	3	**10**
Tot	**3**	**5**	**4**	**8**	**20**

L'obiettivo dell'AC è **studiare la struttura dell'interdipendenza tra le mutabili Genere e Diploma**, considerando le "corrispondenze" tra gli elementi dei due insiemi

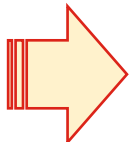
La "domanda chiave" è: **esiste una relazione tra il Genere e il tipo di Diploma?**

TRASFORMAZIONE DEI DATI E LE FREQUENZE RELATIVE

Dalla matrice iniziale, **N**, è possibile passare alla **tabella delle frequenze relative**, **F**, dividendo ciascun valore per il totale:

N=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	1	1	3	5	**10**
M	2	4	1	3	**10**
Tot	**3**	**5**	**4**	**8**	**20**



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

F=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	0,05	0,05	0,15	0,25	**0,5**
M	0,10	0,20	0,05	0,15	**0,5**
Tot	**0,15**	**0,25**	**0,20**	**0,40**	**1,00**

TRASFORMAZIONE DEI DATI E PROFILI RIGA E PROFILI COLONNA

Utilizzando le **distribuzioni condizionate**, si definiscono le tabelle dei **Profili**:

N=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	1	1	3	5	**10**
M	2	4	1	3	**10**
Tot	**3**	**5**	**4**	**8**	**20**

$r_{ij} = \frac{n_{ij}}{n_{i.}} \times 100$

R=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	10	10	30	50	**100**
M	20	40	10	30	**100**
AVG Profile	15	**25**	**20**	**40**	

$c_{ij} = \frac{n_{ij}}{n_{.j}} \times 100$

C=

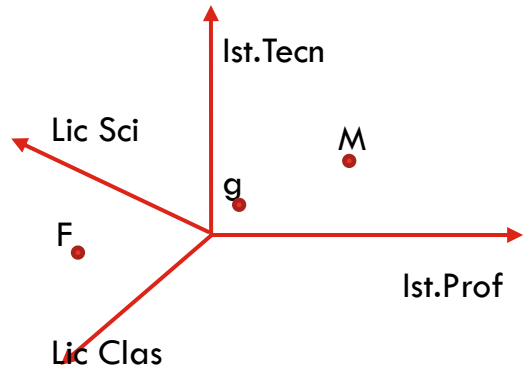
	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Avg Profile
F	33,3	20	75	62,5	**50**
M	66,7	80	25	37,5	**50**
Tot	100	**100**	**100**	**100**	

APPROCCIO GEOMETRICO PROFILI RIGA

Consideriamo la tabella dei **Profili Riga**:
(discorso assolutamente analogo può essere fatto per i Profili Colonna)

R=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	10	10	30	50	**100**
M	20	40	10	30	**100**
AVG Profile	15	**25**	**20**	**40**	



Ognuna delle *I* righe della tabella può essere considerata come **un punto nello spazio a J dimensioni generato dalle colonne della tabella**, con coordinate date dai valori del profilo corrispondente.

A differenza di quanto accade in ACP, dove tutte le righe hanno lo stesso peso, 1/n, nell'Analisi delle Corrispondenze ogni profilo ha un peso definito dal suo marginale relativo, n_i/n.

Il **Profilo riga medio** può essere ottenuto dalla **media ponderata delle J colonne**, con pesi pari alle frequenze relative di ciascuna riga.

APPROCCIO GEOMETRICO PROFILI RIGA

Consideriamo la tabella dei **Profili Riga**:
(discorso assolutamente analogo può essere fatto per i Profili Colonna)

R=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	10	10	30	50	**100**
M	20	40	10	30	**100**
AVG Profile	15	**25**	**20**	**40**	

(10*0,5)+(20*0,5)= 15

N=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot	Weight
F	1	1	3	5	**10**	**0,5**
M	2	4	1	3	**10**	**0,5**
Tot	**3**	**5**	**4**	**8**	20	**1,00**

Il **Profilo riga medio** può essere ottenuto dalla **media ponderata delle J colonne**, con pesi pari alle frequenze relative di ciascuna riga.

APPROCCIO GEOMETRICO PROFILI RIGA

Consideriamo la tabella dei **Profili Riga**:
(discorso assolutamente analogo può essere fatto per i Profili Colonna)

R=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot
F	10	10	30	50	**100**
M	20	40	10	30	**100**
AVG Profile	15	**25**	**20**	**40**	

N=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot	Weight
F	1	1	3	5	**10**	**0,5**
M	2	4	1	3	**10**	**0,5**
Tot	**3**	**5**	**4**	**8**	20	**1,00**

(10*0,5)+(20*0,5)= 15
(10*0,5)+(40*0,5)= 25
(30*0,5)+(10*0,5)= 20
(50*0,5)+(30*0,5)= 40

Il **Profilo riga medio** può essere ottenuto dalla **media ponderata delle J colonne**, con pesi pari alle frequenze relative di ciascuna riga

APPROCCIO GEOMETRICO PROFILI COLONNA

Consideriamo la tabella dei **Profili Colonna**:

C=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Avg Profile
F	33,3	20	75	62,5	**50**
M	66,7	80	25	37,5	**50**
Tot	100	**100**	**100**	**100**	

N=

	Istituto Professionale	Istituto Tecnico	Liceo Classico	Liceo Scientifico	Tot	Weight
F	1	1	3	5	**10**	**0,5**
M	2	4	1	3	**10**	**0,5**
Tot	**3**	**5**	**4**	**8**	20	**1,00**
Weight	**0,15**	**0,25**	**0,20**	**0,40**	**1,00**	

(33,3*0,15)+(20*0,25) +(75*0,20) +(62,5*0,40)= 50
(66,7*0,15)+(80*0,25) +(25*0,20) +(37,5*0,40)= 50

I **pesi** utilizzati per il calcolo del **profilo riga medio** siano anche le **componenti** del vettore contenente i valori del **profilo colonna medio**

Quindi, a differenza di quanto accade in ACP, in cui la **stessa trasformazione iniziale** dei dati (la centratura), si applica sia alle righe sia alle colonne, determinando, però, **conseguenze molto diverse nei due spazi R^p e R^n**, nell'AC **trasformazioni diverse** dei dati iniziali (le tabelle dei profili si ottengono dividendo una per i marginali di riga l'altra per i marginali di colonna) producono **risultati assolutamente identici**.

LA DISTANZA DEL CHI QUADRO E L'INERZIA

La caratteristica dell'AC di operare sulle tabelle dei profili porta ad utilizzare, nel calcolo delle distanze tra punti, una metrica diversa da quella classica euclidea utilizzata in ACP.

La **distanza euclidea** tende a **dare eccessiva importanza alle modalità più frequenti**.

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

Per risolvere questo problema si rende opportuno **ponderare** ciascuna colonna dando **maggior peso** alle modalità che si presentano con **frequenza minore**.

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

Nel calcolare la **distanza tra due punti in R^J**, la **distanza del chi-quadrato** attribuisce il **peso $1/f_{.j}$** ad ogni dimensione j.

LA DISTANZA DEL CHI QUADRO E L'INERZIA

Con la **metrica del chi-quadro** è possibile capire se le variabili sono indipendenti tra di loro oppure associate e studiare **l'inerzia totale** della nuvola dei punti rispetto al **baricentro G**

Infatti se tutti i profili sono uguali tra loro, e uguali al profilo medio, allora c'è **indipendenza** tra le variabili. La nube è, di fatto, un singolo punto, nel baricentro.
Se le variabili sono **associate**, più lontani sono i dati dalla situazione di indipendenza, più dispersi attorno al baricentro saranno i profili.

L'inerzia è dunque un indicatore della **dispersione della nube dei profili attorno al baricentro G**, ed è una **misura dell'intensità del legame tra i caratteri osservati**. Esaminare la dispersione dei punti attorno al baricentro G equivale, dunque, a esaminare lo scarto tra i dati e il modello di indipendenza.
L'inerzia della nube dei profili è calcolata a partire dai valori della **tabella di contingenza**. È, questa, una importante **differenza** rispetto all'**ACP** normalizzata, in cui l'inerzia totale della nube degli individui risulta pari al numero delle variabili, p, e dipende, dunque, dalle dimensioni della matrice dei dati e non dai dati stessi.

LA RAPPRESENTAZIONE DELLE NUBI N_I E N_J

Come in ACP, l'origine degli assi coincide con il baricentro G_I della nube N_I nello spazio $\mathbb{R}^J$.

L'obiettivo è individuare una sequenza di **fattori ortogonali**, di **inerzia massima**, sui quali **proiettare i profili**.

Tutto quanto detto per **la nube N_I nello spazio $\mathbb{R}^J$** , vale, **simmetricamente** per **la nube N_J nello spazio $\mathbb{R}^I$** .

Inoltre, ricordiamo che:

- L'**inerzia totale** misura l'intensità della relazione tra le due variabili *V1* e *V2* della tabella di contingenza.
  L'**autovalore** misura la parte di inerzia associata alla dimensione; dipende dalle coordinate dei profili riga proiettati sulla singola dimensione, e misura la quota di variabilità spiegata da quella dimensione.

- La **distanza di un profilo dall'origine** può essere interpretata come **distanza dal profilo medio** e, quindi, contribuisce a spiegare la relazione tra *V1* e *V2* .

- La **prossimità tra due profili**, i e i', esprime un analogo allontanamento dal profilo medio. Queste modalità della variabile *V1* sono generalmente associate alle stesse modalità della variabile *V2*.

- Il fatto che due profili, , i e i', siano **opposti rispetto all'origine** può essere interpretato come **due modi opposti di differenziarsi rispetto al profilo medio**; le modalità di *V2* alle quali il profilo i si associa maggiormente, sono anche quelle alle quali il profilo *i'* si associa di meno, avendo sempre come riferimento la situazione di indipendenza. possibile

LA RELAZIONE DUALE E RAPPRESENTAZIONE CONGIUNTA

Le analisi delle nubi dei profili riga e dei profili colonna sono strettamente legate da quelle che vengono definite le **relazioni di dualità**, che si spiegano ricordando che le analisi sono condotte sulla **stessa tabella** ma **da due punti di vista** (quello delle righe e quello delle colonne) e che le righe e le colonne di una tabella di contingenza sono, intrinsecamente, oggetti della stessa natura, e cioè modalità di variabili qualitative.

Indicando con:

$F_s(i)$: la coordinata del profilo riga i sull'asse di rango s (in $\mathbf{R}^J$);
$G_s(j)$: la coordinata del profilo colonna j sull'asse di rango s (in $\mathbf{R}^I$);
λ_s : l'inerzia della nube N_I (o N_J) proiettata sull'asse di rango s in $\mathbf{R}^J$ (o in $\mathbf{R}^I$)

valgono le seguenti importanti relazioni:

$$F_s(i) = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} \cdot G_s(j) \qquad G_s(j) = \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} \cdot F_s(i)$$

Quindi, **ciascuna riga i è baricentro dei vertici-colonna**, con pesi definiti dai valori del profilo riga corrispondente, così come **ciascuna colonna j è baricentro dei vertici-riga**, con pesi definiti dai valori del profilo colonna corrispondente.

LA RELAZIONE DUALE E RAPPRESENTAZIONE CONGIUNTA

$$F_s(i) = \sum_{j=1}^{J} \frac{f_{ij}}{f_{i.}} \cdot G_s(j) \qquad G_s(j) = \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} \cdot F_s(i)$$

Nell'AC, **ciascuna riga i è baricentro dei vertici-colonna**, con pesi definiti dai valori del profilo riga corrispondente,
ciascuna colonna j è baricentro dei vertici-riga, con pesi definiti dai valori del profilo colonna corrispondente.

Questa proprietà non consente, direttamente, la rappresentazione simultanea, poiché ogni punto di uno spazio risulta baricentro dei punti dell'altro spazio, ciascuno dei quali, a sua volta, risulta baricentro dei punti del primo

Per risolvere questo problema, ogni profilo viene fatto "esplodere", moltiplicando la coordinata per che, per costruzione, è maggiore di 1 (essendo ogni autovalore compreso tra zero e uno).

$$\frac{1}{\sqrt{\lambda}}$$

Con la **rappresentazione simultanea**, **l'AC definisce in modo chiaro e semplice le caratteristiche della tabella analizzata**, caratteristiche che non sono sempre evidenti quando si osserva la tabella dei dati.

SOMIGLIANZE E DIFFERENZE TRA PUNTI-MODALITA'

L'AC elabora le frequenze delle modalità delle due variabili considerate e le trasforma in misure di distanza che permettono di valutare le somiglianze e le differenze fra le righe e fra le colonne di una tabella di contingenza.

Il termine **CORRISPONDENZE** sottolinea il fatto che l'analisi indica quali cella della tabella forniscono il maggior contributo al fattore.

L'attrazione fra la i-esima modalità della variabile X e la j-esima modalità della variabile Y si traduce in una vicinanza fra la corrispondente coppia di punti sull'asse fattoriale.

Le più forti opposizioni fra coppie di modalità di due variabili diverse sono da intendersi come corrispondenti distanze fra i relativi punti-modalità sul piano fattoriale

Una forte vicinanza su un piano fattoriale fra due modalità della stessa variabile significa che i corrispondenti profili presentano una struttura molto simile, mentre una loro lontananza (in termini di opposizione rispetto all'origine degli assi) sta ad indicare che questi profili sono opposti.

L'INTERPRETAZIONE DEI FATTORI

CONTRIBUTO ASSOLUTO di ciascuna modalità attiva

Rappresenta la parte di inerzia del fattore dovuta alla modalità cui si riferisce. Con il contributo assoluto si valuta **quanta influenza una modalità ha avuto nel determinare un certo fattore**, in rapporto all'insieme delle modalità attive

COSENO QUADRATO (CONTRIBUTO RELATIVO)

È un valore che permette di valutare il contributo che un certo fattore F fornisce alla riproduzione della inerzia di ogni modalità attiva.
Se questo contributo è basso la modalità in questione non è ben rappresenta sul fattore F e la sua dispersione sarà riprodotta da altre dimensioni.
Al contrario, se il contributo è elevato, è lecito analizzare il ruolo che essa gioca nella formazione dell'asse sul quale è ben rappresentata.
Il coseno quadrato varia da 0 e 1.

COORDINATE FATTORIALI

Possono avere segno positivo o negativo e stabiliscono la posizione delle modalità (sia attive sia illustrative) sugli assi, sia in termini di distanza dal centro-origine, sia in termini di versante (positivo o negativo) dell'asse fattoriale considerato.
Il valore zero corrisponde alla media delle coordinate su un fattore.
Di solito, le modalità attive caratterizzate da valori alti nelle coordinate fattoriali sono quelli che contribuiscono di più alla formazione dell'asse stesso.

I DATI: ASPETTI DI VITA QUOTIDIANA

PERIODO DI RIFERIMENTO: **ANNI 2020-2022**

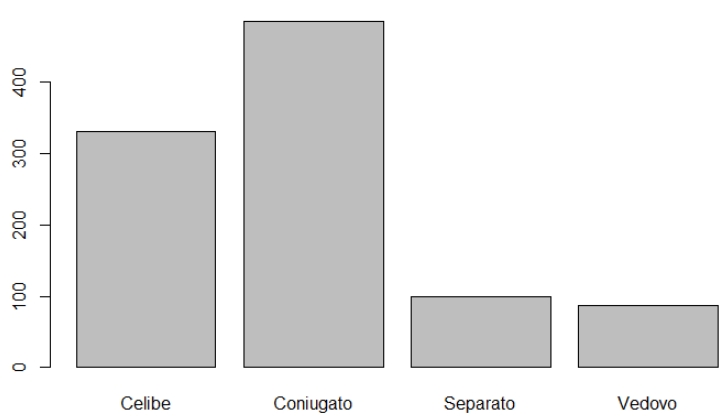
L'indagine campionaria **"Aspetti della vita quotidiana"** fa parte di un sistema integrato di indagini sociali – le indagini multiscopo sulle famiglie - e rileva informazioni fondamentali relative alla **vita quotidiana degli individui e delle famiglie**.

A partire dal 1993, l'indagine viene svolta ogni anno. Le informazioni raccolte consentono di conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno e se sono soddisfatti del funzionamento di quei servizi di pubblica utilità che dovrebbero contribuire al miglioramento della qualità della vita. Scuola, lavoro, vita familiare e di relazione, abitazione e zona in cui si vive, tempo libero, partecipazione politica e sociale, salute, stili di vita sono i temi indagati

I DATI: ASPETTI DI VITA QUOTIDIANA

	id	StatoCivile	TempoLIbero
1	1	Coniugato	Poco
2	6	Coniugato	Abbastanza
3	7	Celibe	Molto
4	8	Separato	Abbastanza
5	9	Celibe	Poco
6	10	Celibe	Poco
7	11	Coniugato	Poco
8	12	Coniugato	Abbastanza
9	13	Celibe	Poco
10	14	Celibe	Abbastanza
11	15	Coniugato	Molto
12	16	Coniugato	Abbastanza
13	17	Celibe	Molto

STATISTICHE DESCRITTIVE



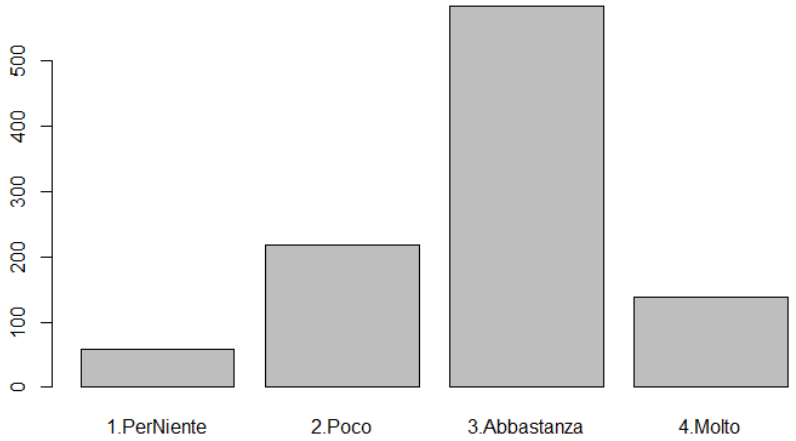


TABELLA DI CONTINGENZA E TEST DEL CHI-QUADRO

 # Costruiamo una tabella di contingenza

```
          1.PerNiente 2.Poco 3.Abbastanza 4.Molto
Celibe            5.2   20.0          57.0    17.9
Coniugato         4.5   24.0          60.3    11.2
Separato          6.1   19.2          58.6    16.2
Vedovo           14.9   20.7          52.9    11.5
>
```

 # Test del Chi-quadro

```
> chisq

        Pearson's Chi-squared test

data:  x
X-squared = 23.965, df = 9, p-value = 0.004357

>
```

le variabili riga e colonna sono associate in modo statisticamente significativo (p-value = 0.0041)

GLI AUTOVALORI

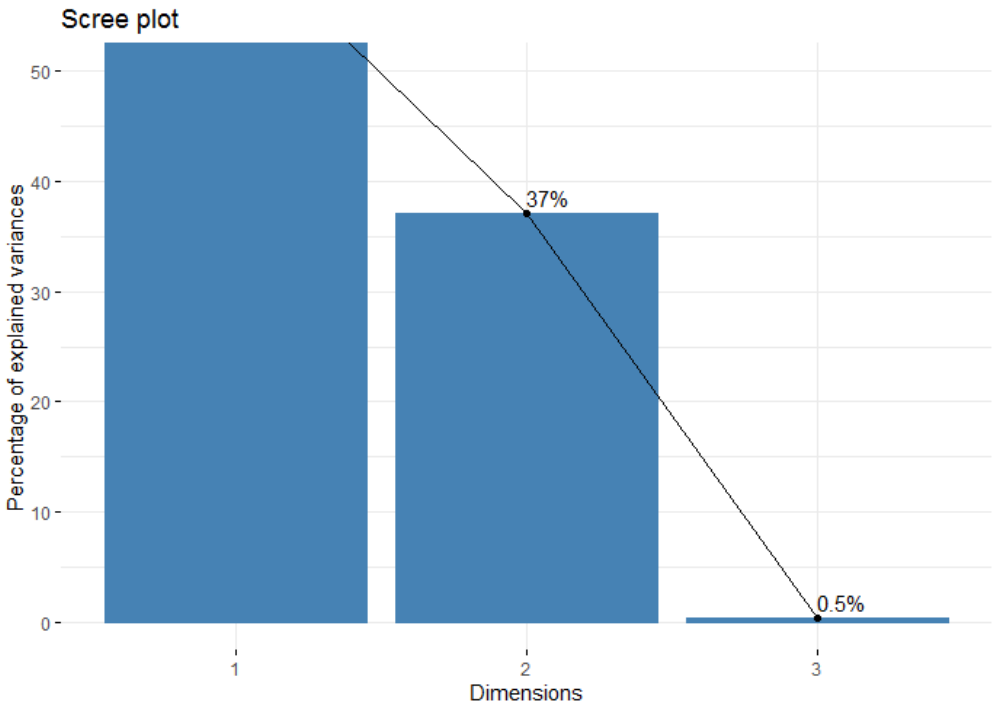
```
# compute CA

eig.val <- get_eigenvalue(res.ca)
```

La prima informazione che troviamo, dopo il chi quadrato della tabella, sono gli autovalori (eigenvalues), che corrispondono alla quantità di informazioni trattenute da ciascun asse. Gli autovalori possono essere utilizzati per determinare il numero di assi da conservare.

```
> round(eig.val, digits=3)
      eigenvalue variance.percent cumulative.variance.percent
Dim.1      0.015           62.494                       62.494
Dim.2      0.009           37.045                       99.539
Dim.3      0.000            0.461                      100.000
>
```

Le tre dimensioni riproducono il 100% della variabilità totale, mentre le prime due raggiungono l'99,5%. Consideriamo solo le prime due dimensioni.

GLI AUTOVALORI

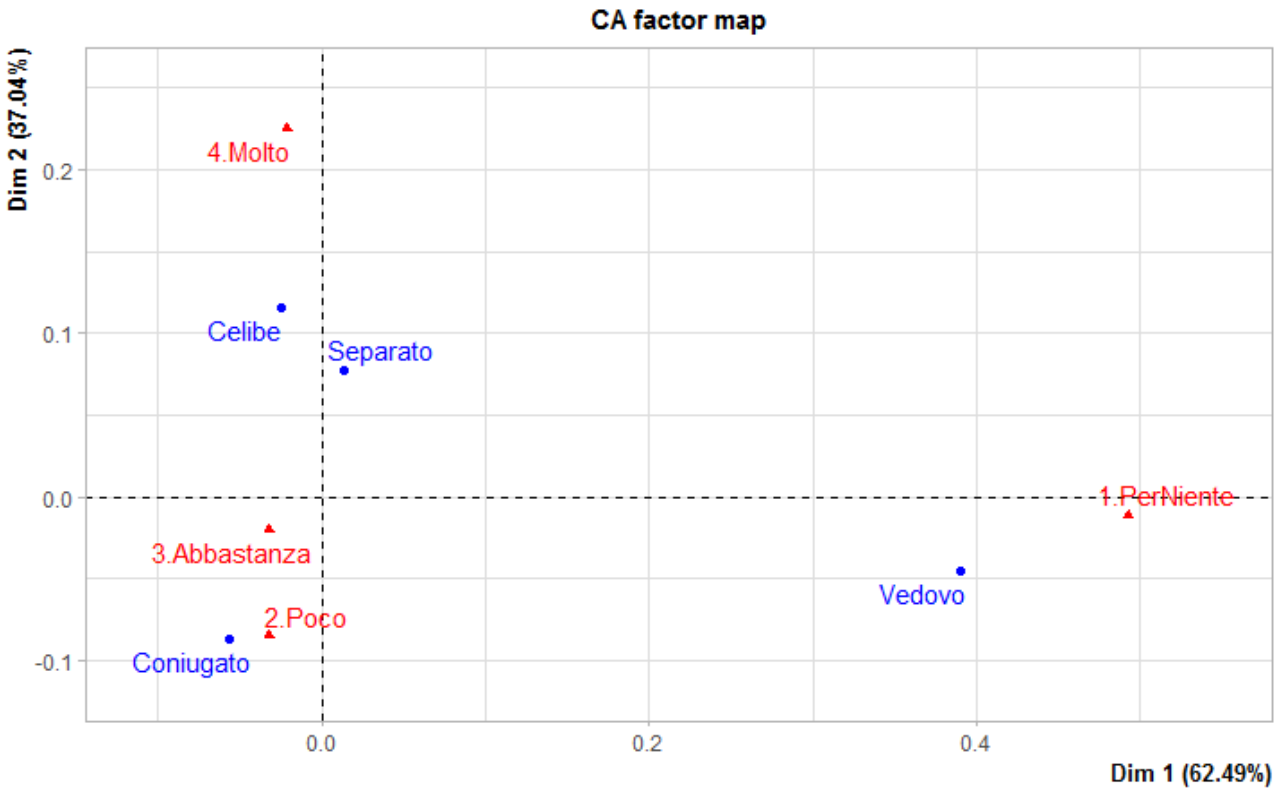


Un metodo alternativo per determinare il numero di dimensioni è guardare uno Scree Plot, che è il grafico degli autovalori/varianze ordinate dal più grande al più piccolo.

Il numero di componenti è determinato nel punto, oltre il quale gli autovalori rimanenti sono tutti relativamente piccoli e di dimensioni comparabili.

Secondo il grafico, dovrebbero essere utilizzate solo le dimensioni 1 e 2. La dimensione 3 spiega solo l0 0,5% dell'inerzia totale

IL GRAFICO E LE COORDINATE

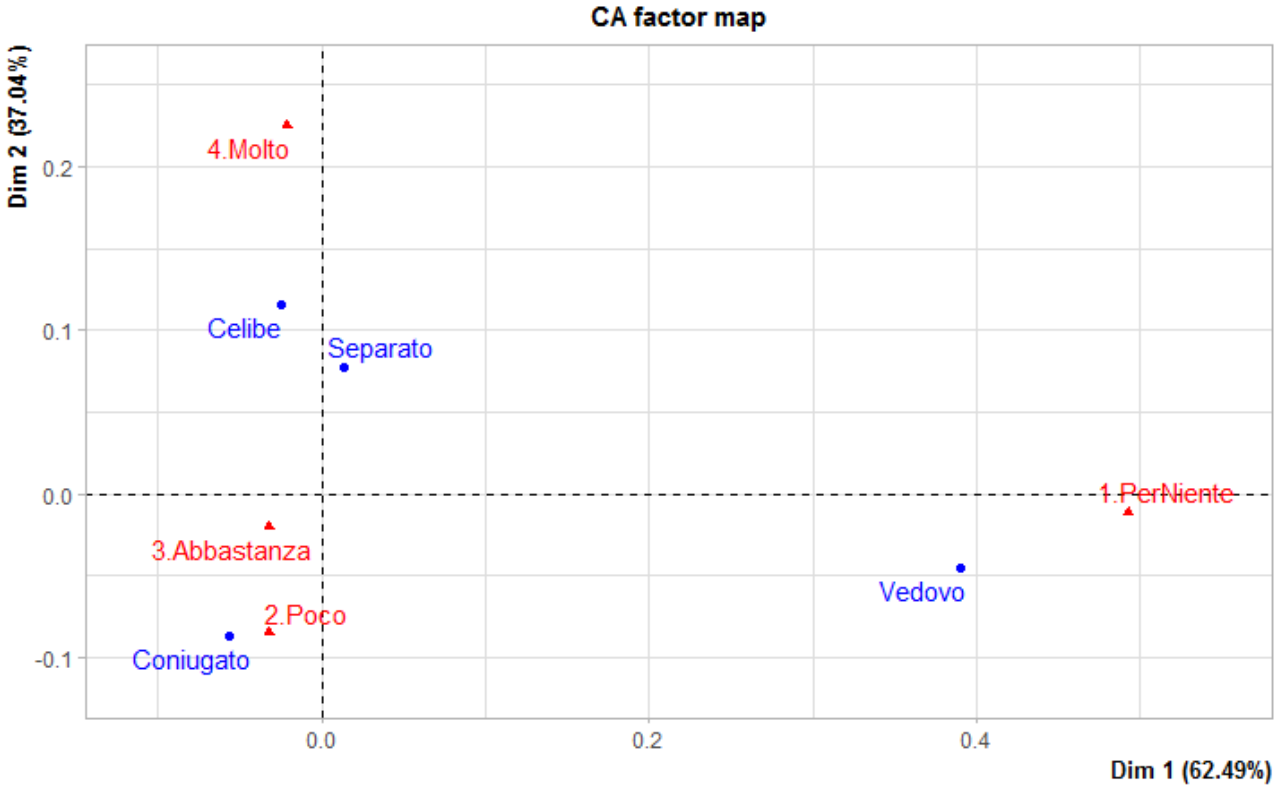


Il grafico sopra è chiamato grafico simmetrico e mostra un modello globale all'interno dei dati.
Le righe sono rappresentate da punti blu e le colonne da triangoli rossi.
La distanza tra qualsiasi punto di riga o punto di colonna fornisce una misura della loro somiglianza (o dissomiglianza).
I punti riga con profilo simile vengono rappresentati vicini sulla mappa dei fattori. Lo stesso vale per i punti colonna.

L'origine del grafico rappresenta il profilo medio di riga e di colonna (indipendenza statistica); le modalità più distanti dall'origine sono quelle più discriminanti o caratteristiche (analogamente al **test di chi quadrato**).

Sulla prima dimensione a destra troviamo *Vedovo*
Sulla seconda dimensione, in alto sono rappresentate *Celibe* e *Separato*, mentre in basso troviamo *Coniugato*

IL GRAFICO E LE COORDINATE



Il **grafico simmetrico** rappresenta i profili di riga e colonna contemporaneamente in uno spazio comune. In questo caso, solo la distanza tra i punti di riga o la distanza tra i punti di colonna può essere realmente interpretata.

La distanza tra gli elementi riga e colonna non è significativa! Possiamo solo fare affermazioni generali sul modello osservato.

Per interpretare la distanza tra i punti colonna e riga, i profili colonna devono essere presentati nello spazio riga o viceversa. Questo tipo di mappa è chiamato **biplot asimmetrico**

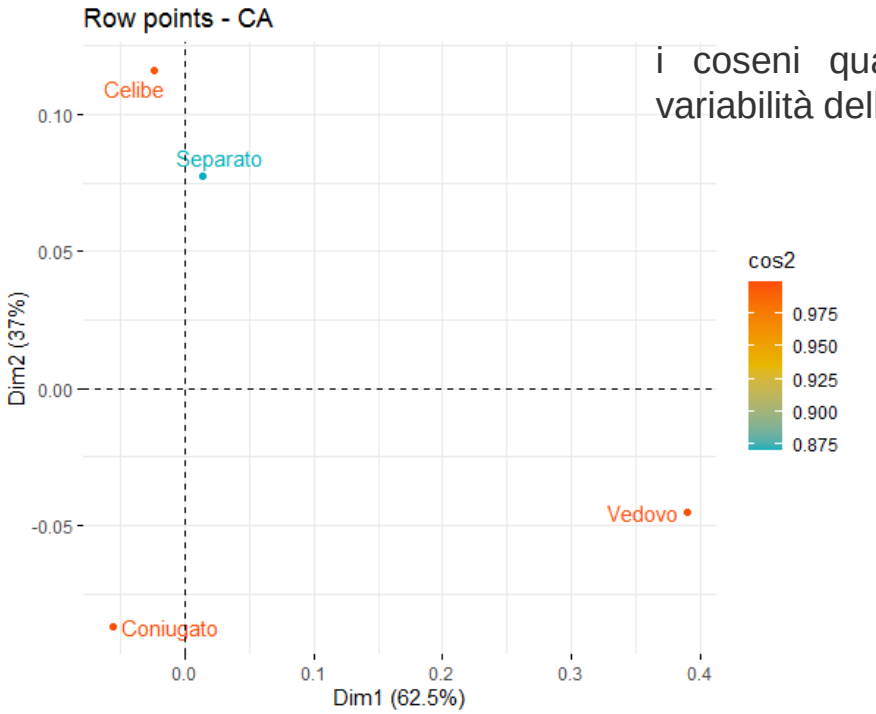
ANALISI DELLE RIGHE

```
>
> row$coord
                Dim 1       Dim 2         Dim 3
Celibe    -0.02428862  0.11606271  0.0072320768
Coniugato -0.05624234 -0.08690119  0.0006856007
Separato   0.01318286  0.07777600 -0.0304734033
Vedovo     0.39001644 -0.04529127  0.0034304694
> row$cos2
                Dim 1      Dim 2        Dim 3
Celibe    0.04180157 0.95449236 3.706066e-03
Coniugato 0.29519892 0.70475722 4.386626e-05
Separato  0.02430082 0.84584875 1.298504e-01
Vedovo    0.98661874 0.01330493 7.632919e-05
> row$contrib
                Dim 1     Dim 2     Dim 3
Celibe     1.2998812 50.072530 15.627588
Coniugato 10.2224797 41.171533  0.205987
Separato   0.1148785  6.745694 83.239428
Vedovo    88.3627606  2.010242  0.926997
>
```

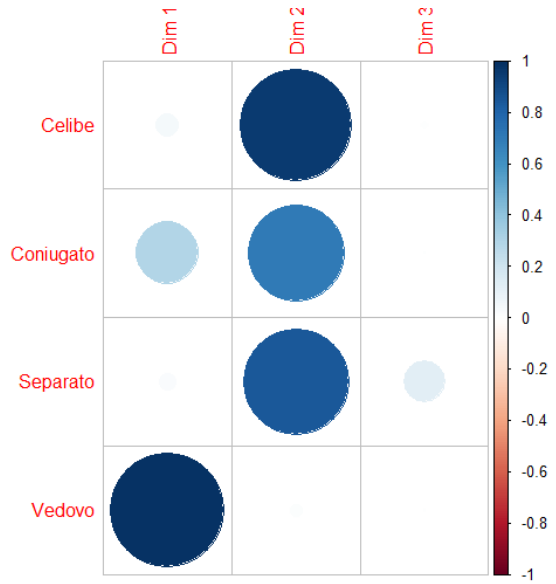
I **contributi** indicano il contributo dato dalla modalità alla costruzione dell'asse;
i **coseni quadrati** indicano la quota di variabilità della modalità riprodotta dall'asse.

La somma per colonna dei contributi è pari a 1, quindi la soglia di significatività del contributo di una modalità è pari a 1 / n.modalità (in questo caso, 1/4 per le modalità di riga e 1/4 per quelle di colonna).
La somma dei coseni quadrati è invece pari a 1 per riga.

ANALISI DELLE RIGHE – I COSENI QUADRATI



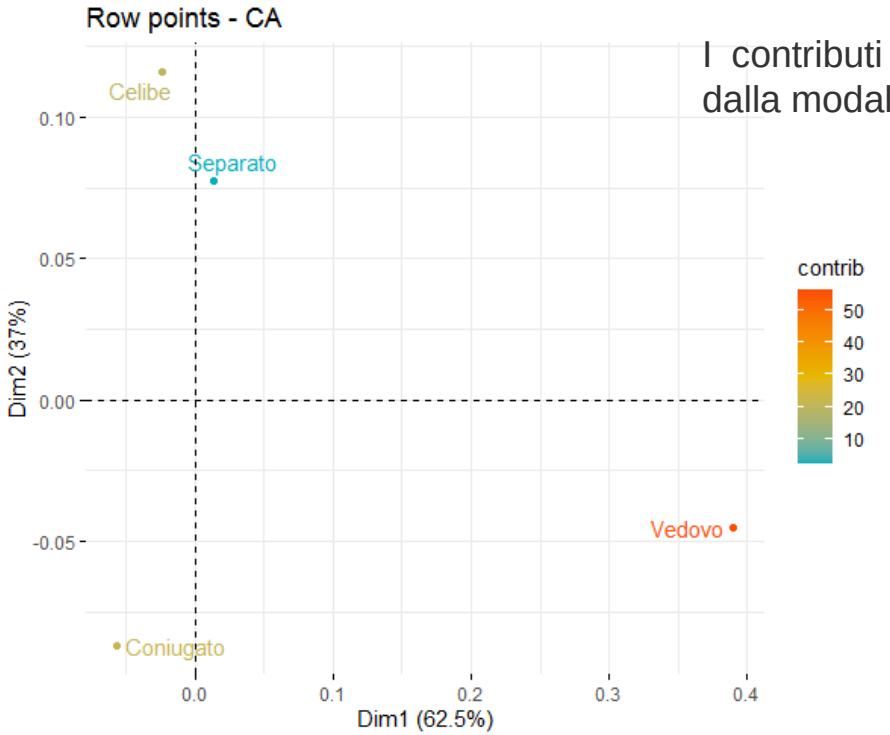
i coseni quadrati indicano la quota di variabilità della modalità riprodotta dall'asse



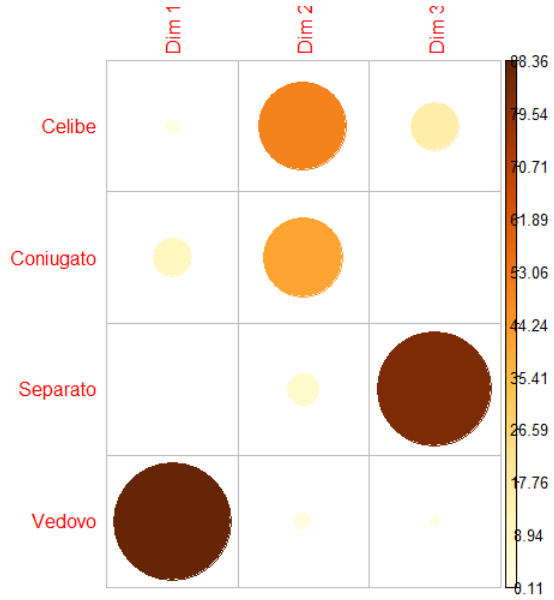
Il grafico mostra le relazioni tra i punti di riga:

- Le righe con un profilo simile vengono raggruppate insieme.
- Le righe correlate negativamente sono posizionate sui lati opposti dell'origine del grafico (quadranti opposti).
- La distanza tra i punti di riga e l'origine misura la qualità dei punti di riga sulla mappa dei fattori. I punti di riga che sono lontani dall'origine sono ben rappresentati sulla mappa dei fattori.

ANALISI DELLE RIGHE – I CONTRIBUTI



I contributi indicano il contributo dato dalla modalità alla costruzione dell'asse

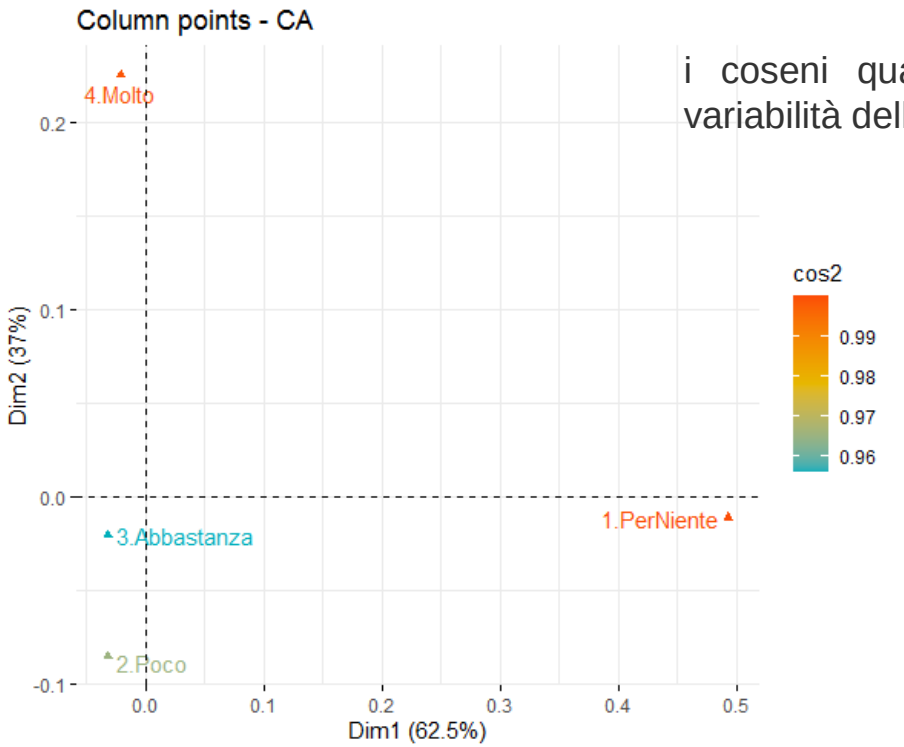


Il grafico a dispersione dà un'idea di quale polo delle dimensioni stiano effettivamente contribuendo le categorie di riga.
È evidente che la categoria Vedovo ha un contributo importante al polo positivo della prima dimensione, mentre le categorie Celibe e Coniugato hanno un contributo importante al polo positivo e negativo della seconda dimensione

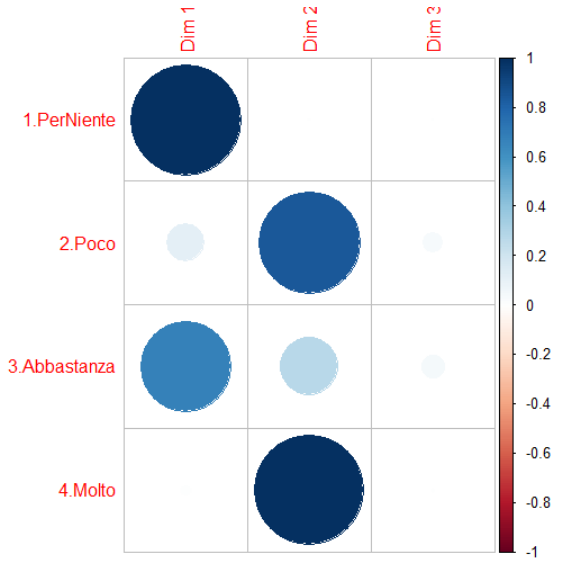
ANALISI DELLE COLONNE

```
> col$coord
                 Dim 1       Dim 2         Dim 3
1.PerNiente   0.49296617 -0.01114225 -0.0003540322
2.Poco       -0.03195356 -0.08519219  0.0172061206
3.Abbastanza -0.03195292 -0.02049509 -0.0081190199
4.Molto      -0.02110576  0.22498183  0.0071503676
> col$cos2
                 Dim 1        Dim 2        Dim 3
1.PerNiente  0.999488874 0.0005106106 5.155008e-07
2.Poco       0.119073453 0.8464008496 3.452570e-02
3.Abbastanza 0.677517376 0.2787398374 4.374279e-02
4.Molto      0.008714999 0.9902847201 1.000281e-03
> col$contrib
                Dim 1       Dim 2        Dim 3
1.PerNiente  94.1123079  0.08110997  0.006582117
2.Poco        1.4930238 17.90374477 58.703231452
3.Abbastanza  3.9812391  2.76320352 34.855557384
4.Molto       0.4134292 79.25194173  6.434629047
>
```

ANALISI DELLE COLONNE – I COSENI QUADRATI

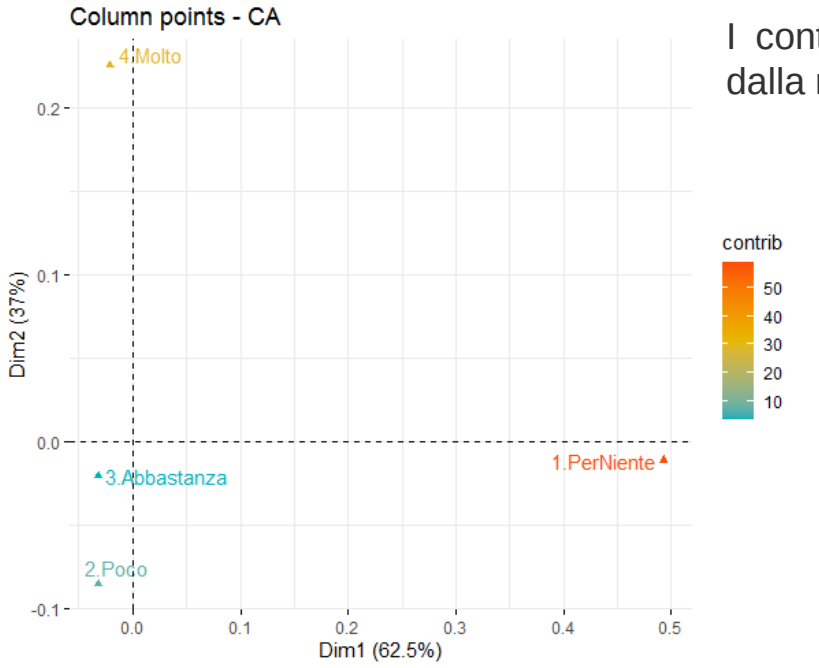


i coseni quadrati indicano la quota di variabilità della modalità riprodotta dall'asse

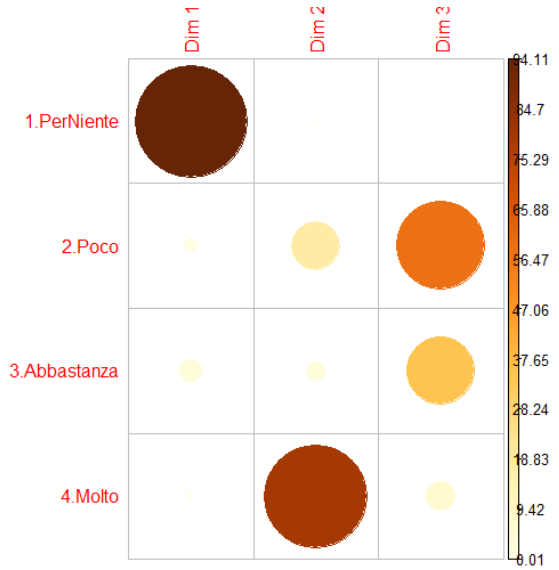


Per Niente e *Molto* risultano lontani dall'origine e quindi sono ben rappresentati sulla mappa dei fattori.

ANALISI DELLE COLONNE – I CONTRIBUTI

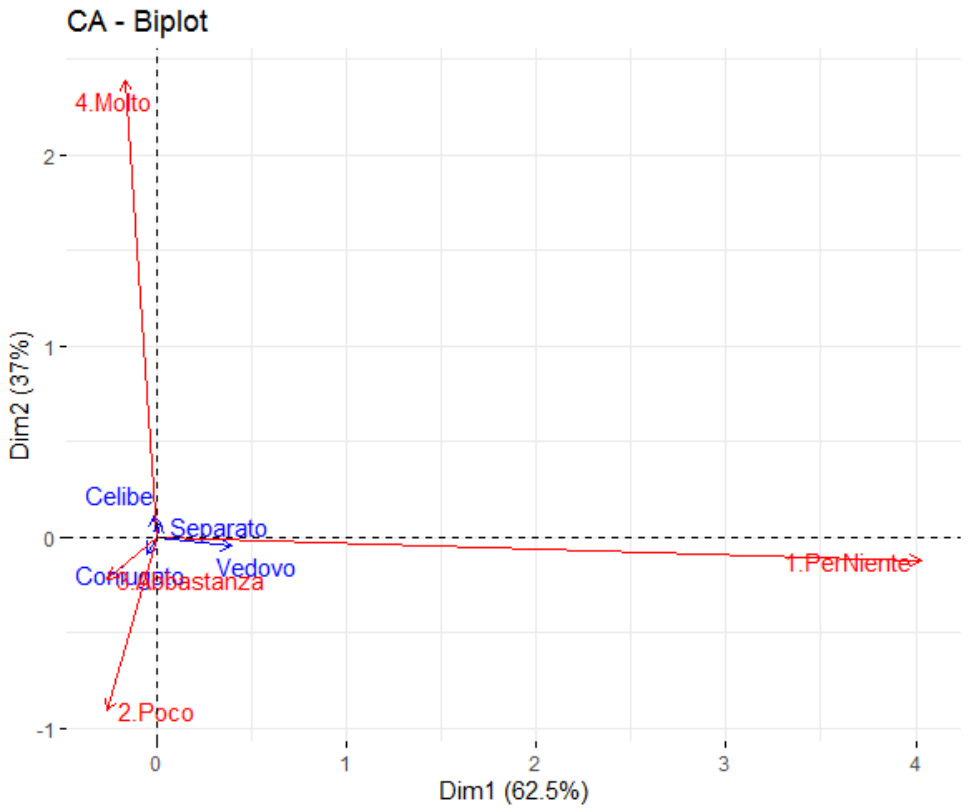


I contributi indicano il contributo dato dalla modalità alla costruzione dell'asse



È evidente che la categoria *Per Niente* ha un contributo importante al polo positivo della prima dimensione, mentre la categoria *Molto* ha un contributo importante al polo positivo della seconda dimensione

IL GRAFICO ASIMMETRICO



Nel grafico, la posizione dei punti del profilo della colonna è invariata rispetto a quella del biplot convenzionale. Tuttavia, le distanze dei punti di riga dall'origine del grafico sono correlate ai loro contributi alla mappa fattoriale bidimensionale.
Più una freccia è vicina (in termini di distanza angolare) a un asse, maggiore è il contributo della categoria di riga su quell'asse rispetto all'altro asse.
Se la freccia è a metà strada tra i due, la sua categoria di riga contribuisce ai due assi nella stessa misura.

È evidente che la categoria Per Niente ha un contributo importante al polo positivo della prima dimensione, mentre la categoria Abbastanza ha un contributo relativamente importante al polo negativo della prima dimensione
La dimensione 2 è definita principalmente dalla categoria di riga Molto (polo positivo) e Poco (polo negativo).


ORA TOCCA
A VOI!

LE FACCENDE DOMESTICHE

 # Libraries

```
library(FactoMineR)
library(factoextra)
```

 # Data Import

```
library(LabRS)
data("faccende")
```

	Moglie	Alternati	Marito	Insieme
Bucato	156	14	2	4
Pasto_princ	124	20	5	4
Cena	77	11	7	13
Colazione	82	36	15	7
Rassettare	53	11	1	57
Piatti	32	24	4	53
Shopping	33	23	9	55
Documenti	12	46	23	15
Guidare	10	51	75	3
Finanze	13	13	21	66
Assicurazioni	8	1	53	77
Riparazioni	0	3	160	2
Ferie	0	1	6	153

SUMMARY

 # compute CA

res.ca <- CA(faccende)

```
> summary(res.ca)

Call:
CA(X = faccende)

The chi square of independence between the two variables is equal to 1944.456 (p-value =  0 ).

Eigenvalues
                       Dim.1   Dim.2   Dim.3
Variance               0.543   0.445   0.127
% of var.             48.692  39.913  11.395
Cumulative % of var.  48.692  88.605 100.000

Rows (the 10 first)
                | Iner*1000 |    Dim.1 |     ctr |    cos2 |    Dim.2 |     ctr |    cos2 |    Dim.3 |     ctr |    cos2 |
Bucato          |   134.160 |   -0.992 |  18.287 |   0.740 |    0.495 |   5.564 |   0.185 |   -0.317 |   7.968 |   0.075 |
Pasto_princ     |    90.692 |   -0.876 |  12.389 |   0.742 |    0.490 |   4.736 |   0.232 |   -0.164 |   1.859 |   0.026 |
Cena            |    38.246 |   -0.693 |   5.471 |   0.777 |    0.308 |   1.321 |   0.154 |   -0.207 |   2.097 |   0.070 |
Colazione       |    41.124 |   -0.509 |   3.825 |   0.505 |    0.453 |   3.699 |   0.400 |    0.220 |   3.069 |   0.095 |
Rassettare      |    24.667 |   -0.394 |   1.998 |   0.440 |   -0.434 |   2.966 |   0.535 |   -0.094 |   0.489 |   0.025 |
Piatti          |    19.587 |   -0.189 |   0.426 |   0.118 |   -0.442 |   2.844 |   0.646 |    0.267 |   3.634 |   0.236 |
Shopping        |    14.970 |   -0.118 |   0.176 |   0.064 |   -0.403 |   2.515 |   0.748 |    0.203 |   2.223 |   0.189 |
Documenti       |    53.300 |    0.227 |   0.521 |   0.053 |    0.254 |   0.796 |   0.066 |    0.923 |  36.940 |   0.881 |
Guidare         |   101.509 |    0.742 |   8.078 |   0.432 |    0.653 |   7.647 |   0.335 |    0.544 |  18.596 |   0.233 |
Finanze         |    29.564 |    0.271 |   0.875 |   0.161 |   -0.618 |   5.559 |   0.837 |    0.035 |   0.062 |   0.003 |

Columns
                | Iner*1000 |    Dim.1 |     ctr |    cos2 |    Dim.2 |     ctr |    cos2 |    Dim.3 |     ctr |    cos2 |
Moglie          |   301.019 |   -0.838 |  44.462 |   0.802 |    0.365 |  10.312 |   0.152 |   -0.200 |  10.822 |   0.046 |
Alternati       |   117.824 |   -0.062 |   0.104 |   0.005 |    0.292 |   2.783 |   0.105 |    0.849 |  82.549 |   0.890 |
Marito          |   381.373 |    1.161 |  54.234 |   0.772 |    0.602 |  17.787 |   0.208 |   -0.189 |   6.133 |   0.020 |
Insieme         |   314.725 |    0.149 |   1.200 |   0.021 |   -1.027 |  69.118 |   0.977 |   -0.046 |   0.495 |   0.002 |
>
```

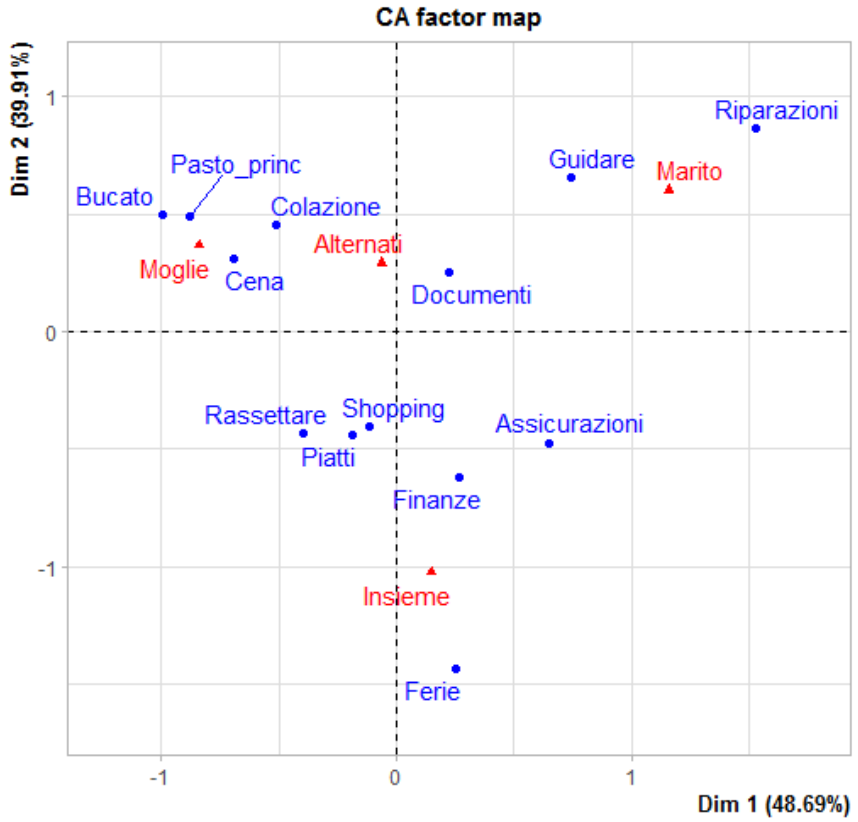
IL GRAFICO

Nel primo quadrante troviamo *Guidare* e *Riparazioni*, vicine (e dunque associate) alla modalità *Marito*

Nel secondo quadrante sono rappresentate *Pasto principale*, *Colazione*, *Bucato* e *Cena*, associate alla modalità *Moglie*.

Al centro rispetto alla prima dimensione, ma non alla seconda, le attività che vengono svolte in maniera *Alternata* (in alto) e *Insieme* (in basso).

Il primo asse (orizzontale) riproduce dunque le differenze legate alla divisione di genere delle attività domestiche; il secondo asse (verticale) distingue invece fra attività di genere e attività svolte assieme o alternandosi.



GLI AUTOVALORI

```
# compute CA

eig.val <- get_eigenvalue(res.ca)
```

La prima informazione che troviamo, dopo il chi quadrato della tabella, sono gli autovalori (eigenvalues), che corrispondono alla quantità di informazioni trattenute da ciascun asse. Gli autovalori possono essere utilizzati per determinare il numero di assi da conservare.

```
>
> get_eigenvalue(res.ca)  # tabella degli autovalori
      eigenvalue variance.percent cumulative.variance.percent
Dim.1  0.5428893         48.69222                    48.69222
Dim.2  0.4450028         39.91269                    88.60491
Dim.3  0.1270484         11.39509                   100.00000
>
```

Le tre dimensioni riproducono il 100% della variabilità totale, mentre le prime due raggiungono l'88,6%. Consideriamo solo le prime due dimensioni.

Prof.ssa Rosanna Cataldo