ANALISI FATTORIALE DELLE CORRISPONDENZE (AFC)

DOTT. MARIO CANDELA

INTRODUZIONE

L'**analisi multivariata** di tabelle caratterizzate da variabili di tipo categorico ha avuto un'evoluzione più lenta rispetto a quanto avvenuto per le variabili numeriche.
Il contributo di studiosi come F. Galton, R.A. Fisher, G.U. Yule allo sviluppo dei metodi di Regressione su variabili continue ed il parallelo affermarsi di metodi quali l'Analisi della Varianza hanno dato per diverso tempo dato alla ricerca statistica carattere di tipo sperimentale, da applicare a variabili misurabili in termini quantitativi, sacrificando quelle non numeriche; difatti, l'interesse per variabili qualitative si risolveva nel tentativo di adattare a queste ultime metodi nati per variabili continue.
La varietà di indici proposti e le interpretazioni del concetto di associazione tra variabili qualitative hanno palesato come non sia possibile definire un unico indice in grado di misurare la *forza* del legame tra 2 caratteri, ma come la scelta di tale misura dipenda dal tipo di dati e dagli obiettivi fissati.

J.P. Benzécri propose un diverso approccio al problema dello studio dell'associazione tra variabili categoriche, in cui non si ipotizza alcuna distribuzione o modello per i dati osservati, la cui matrice viene invece approssimata da una matrice di rango ridotto con lo stesso contenuto di informazione ed in grado di definirne le relazioni mediante la definizione di operatori di proiezione associati a piani fattoriali appropriati.
Tale approccio viene battezzato da Benzécri **Analisi delle Corrispondenze** (**AC**) e rappresenta il punto di partenza di quello che sarebbe diventata la scuola francese di *Analyse des Données;* inoltre, assurge l'aspetto geometrico e le relative proprietà, mista a nessuna formulazione di ipotesi *a priori* sulle configurazioni dei dati.

L'**Analisi delle Corrispondenze** costituisce uno dei più noti ed efficaci strumenti per il trattamento multidimensionale di dati qualitativi.
Se i principi di tale analisi possono riferirsi a Guttman e Fisher, risulta pacifico attribuirne a Benzécri la formalizzazione.
All'inizio l'analisi delle corrispondenze era riferita allo studio delle relazioni esistenti tra gli elementi di 2 insiemi rappresentati dalle modalità di 2 caratteri, riportate sulle righe e sulle colonne di una tabella di contingenza (**Analisi delle Corrispondenze Binarie**).
In seguito, il metodo si è esteso al caso di più variabili qualitative (**Analisi delle Corrispondenze Multiple**).

L'**AC** opera su **tabelle di contingenza** e fornisce, delle descrizioni di tabelle che presentano una codifica disgiuntiva, trovandoci di fronte a tabelle, talvolta definite **di dipendenza** o **incrociate**, in cui ad ogni coppia di numeri ($\boldsymbol{i}$, $\boldsymbol{j}$) corrisponde un numero positivo $\boldsymbol{n_{ij}}$ che è una frequenza.
L'aggettivo *binarie* che accompagna tale analisi si riferisce al fatto che questa consiste nell'incontro di 2 insiemi $\boldsymbol{I}$ e $\boldsymbol{J}$, considerati finiti per definizione, i cui ruoli sono identici anche se la loro origine può risultare diversa.

Ribadiamo che l'Analisi delle Corrispondenze ha come obiettivo quello di individuare dimensioni tese a riassumere l'intreccio di relazioni di *interdipendenza* tra le variabili originarie qualitative; difatti, tramite tale analisi si può trasformare una tabella di contingenza in una rappresentazione grafica, al fine di facilitare l'interpretazione delle informazioni ivi contenute.
Differenza fondamentale tra l'**AC** e l'**ACP**, oltre al dominio di applicazione, riguarda la matrice dei dati utilizzata, poiché nell'**ACP** si usa una matrice di dati con variabili quantitative, mentre nell'**AC** la matrice dei dati è composta da variabili qualitative.
Osserviamo, dunque, la **matrice dei dati** ai fini della nostra analisi.

Passiamo, quindi, a considerare la tabella di contingenza.
L'**AC** opera soprattutto su **tabelle di contingenza**, tuttavia tale analisi è stata applicata anche ad altre tipologie di tabelle purché avessero il requisito dell'*omogeneità*.
Le **tabelle di contingenza** sono un particolare tipo di tabelle a doppia entrata, utilizzate per rappresentare e analizzare le relazioni tra due o più variabili; difatti, a tal scopo osserviamo al loro interno le frequenze congiunte delle variabili.

Tramite la tabella di contingenza l'Analisi delle Corrispondenze vuole studiare la struttura dell'*interdipendenza* tra la variabile $\boldsymbol{X}$ (ad esempio, regioni italiane) e la variabile $\boldsymbol{Y}$ (ad esempio, il tipo di ricerche sul web) analizzando le corrispondenze tra i due insiemi.

La tabella di contingenza può essere letta sia per riga che per colonna, richiedendo però una prima trasformazione dei dati iniziali.

Il totale marginale della $i-esima$ riga lo indichiamo con $\boldsymbol{n_{i.}}$, mentre con $\boldsymbol{n_j}$ indichiamo il marginale della $j-esima$ colonna. Dunque, abbiamo:

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

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

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

Definiamo $\boldsymbol{F}$ la matrice delle frequenze relative che otteniamo dividendo ciascuna frequenza $\boldsymbol{n_{ij}}$ al totale delle frequenze, così dalle quantità ottenute costruiamo le **tabelle dei profili**, rapportando il valore di ogni cella al corrispondente totale di riga o di colonna.

I profili così ottenuti possono essere confrontati tra loro e anche col profilo medio, dato dal rapporto tra il marginale e il totale generale; difatti, il marginale di riga della tabella **F** rappresenterà il profilo riga medio, ed il marginale di colonna rappresenterà invece il profilo colonna medio.

Passiamo a considerare la scelta della **metrica** e la scelta di operare sulle tabelle dei profili porta ad utilizzare una metrica diversa da quella euclidea, qual è la **metrica del chi-quadrato**, che è pari a:

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

Considerando la metrica del chi-quadrato osserviamo la seguente proprietà:

- **Equivalenza distributiva**: se 2 profili riga uguali o proporzionali vengono aggregati in un unico profilo con massa pari alla somma delle masse, la configurazione dei punti in $\boldsymbol{R^c}$ non cambia, né si modificano le distanze tra i profili colonna in $\boldsymbol{R^r}$, ovviamente la medesima proprietà vale per i profili colonna.

  Tale proprietà è importante poiché consente di raggruppare due o più righe o colonne in una sola, riducendo le dimensioni dello spazio di riferimento senza modificare le informazioni di partenza, garantendo l'invariabilità dei risultati indipendentemente da come le variabili sono state originariamente codificate; dunque, con tale proprietà:

  - Non si ha perdita di informazioni se si aggregano modalità;
  - non si ha guadagno di informazioni se si suddividono categorie omogenee.

Consideriamo, dunque, il **test d'indipendenza** tra righe e colonne.

Prima di fare l'analisi che ci permette di ridurre lo spazio $\boldsymbol{R^c}$ in uno spazio $\boldsymbol{R^r}$, è utile stabilire il grado di indipendenza che esiste tra i due caratteri considerati.

Il test che viene, normalmente, adoperato è il **test del chi-quadrato**, basato sulla distribuzione della variabile casuale definita dalla somma dei quadrati delle differenze tra i valori osservati e i valori teorici rapportati ai valori teorici, avendo posto come *ipotesi nulla* l'*indipendenza* tra i caratteri*:*

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

Partendo dalle frequenze della matrice $\boldsymbol{N}$, i valori teorici $\widehat{\boldsymbol{n}}_{\boldsymbol{ij}}$ $(i = 1, \dots, r;\ j = 1, \dots, c)$ saranno dati dal prodotto dei margini $\boldsymbol{n}_{\boldsymbol{i.}}$ e $\boldsymbol{n}_{\boldsymbol{.j}}$ rapportato al totale delle frequenze:

$$\widehat{\boldsymbol{n}}_{\boldsymbol{ij}} = \frac{\boldsymbol{n}_{\boldsymbol{i.}} * \boldsymbol{n}_{\boldsymbol{.j}}}{\boldsymbol{n}}$$

sotto l'ipotesi $\boldsymbol{H_0}$ di indipendenza tra i caratteri la variabile casuale somma segue una distribuzione del chi-quadrato con $\boldsymbol{g = (r-1)\,(c-1)}$ gradi di libertà:

$$\boldsymbol{X} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \frac{\left(\boldsymbol{n}_{\boldsymbol{ij}} - \widehat{\boldsymbol{n}}_{\boldsymbol{ij}}\right)^{\boldsymbol{2}}}{\widehat{\boldsymbol{n}}_{\boldsymbol{ij}}} \sim \boldsymbol{\chi}^{\boldsymbol{2}}{}_{\boldsymbol{(r-1)(c-1)}}$$

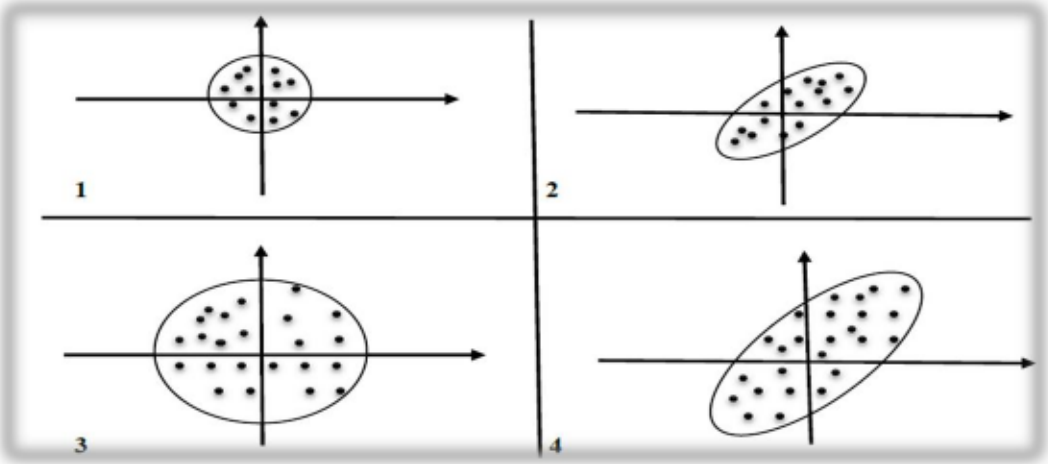
L'ipotesi di indipendenza tra i caratteri sarà rifiutata o meno sulla base del confronto tra il valore ottenuto e il valore teorico riportato sulle tavole, in corrispondenza di un fissato livello di confidenza e dei gradi di libertà.

Dal punto di vista geometrico la situazione di indipendenza può essere rappresentata da una *nube poco dispersa attorno al baricentro* e di forma *sferica*.
Considerando una situazione di interdipendenza possiamo avere diverse forme di nube, in riferimento ai valori assunti dai due indicatori principali:

- **Inerzia totale**, pari alla somma degli *autovalori*, utilizzata per misurare la maggiore o minore dispersione della nube;

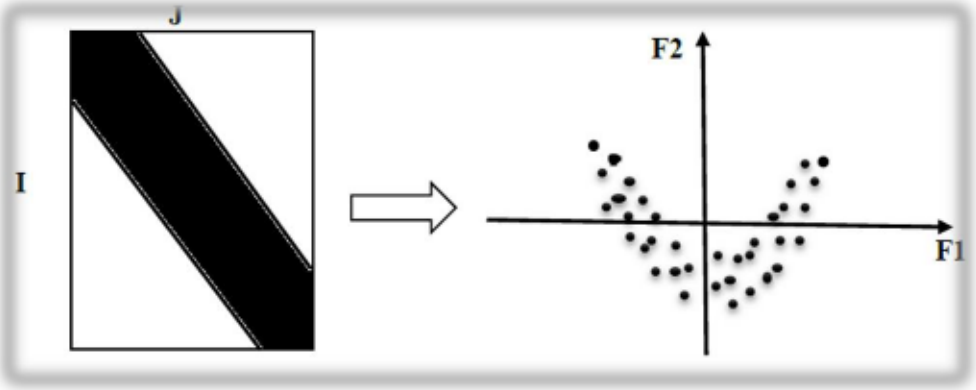
- **Tassi di inerzia**, dati dal rapporto tra i singoli *autovalori* e la loro somma, che identificano la forma della nube.

Secondo quanto detto possono presentarsi quattro situazioni:

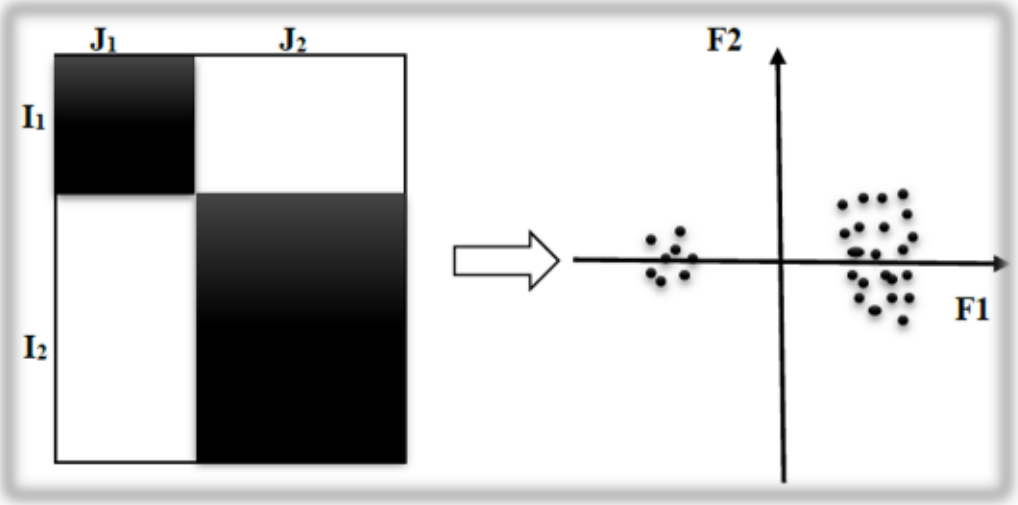


I **casi 1** e **2** rappresentano nubi con uguali tassi di inerzia ma con valori diversi di dispersione totale; difatti, nel 1^o caso abbiamo *autovalori* molto bassi, nel 2^o caso *autovalori* più grandi, ma in entrambi i casi, risultando le direzioni della nube ben definite, l'analisi porterà a risultati interessanti.
I **casi 3** e **4**, invece, rappresentano situazioni con la medesima dispersione totale, ma tassi d'inerzia diversi.
Il caso **3**, pur avendo *autovalori* elevati, ossia da valori del chi-quadrato che porterebbero probabilmente al rifiuto dell'ipotesi di indipendenza, difficilmente produrrà risultati interessanti, data la mancanza di direzioni ben definite; dunque, di aspetti più caratterizzanti.
Il caso **4**, sembrerebbe il **caso ideale**, associando ad una buona dispersione dei punti, ossia valori alti di chi-quadrato, delle ben definite direzioni di allungamento della nube dei punti.

Quindi, la forma della nube dei punti ci dà informazioni sulla relazione tra i due caratteri osservati; inoltre, tale relazione può essere riconosciuta anche osservando la ripartizione delle frequenze nella tabella di contingenza.
Quando si osservano due variabili qualitative ordinate (tipico caso di trasformazioni di variabili quantitative in qualitative mediante suddivisione in classi) la nube assume una forma paraboloide, corrispondente all'**Effetto Guttman** (o **ferro di cavallo**) che rivela una prevalenza di frequenze lungo la diagonale principale della matrice e la ridondanza dell'informazione fornita da una delle variabili che potrebbe essere prevista dalla conoscenza dell'altra.
La struttura dei dati viene rappresentata dal 1^o fattore che risulta dominante mentre i fattori successivi risultano funzioni di ordine superiore al 1^o.



Un altro esempio di disposizione tipica dei punti è quella in cui vi sono due (o più) gruppi ben distinti, che riproduce una **matrice a blocchi**



Ritornando alla tabella osservata del test d'indipendenza tra le righe e le colonne, notiamo il $chi - quadro\ osservato$ e il $p - value$.

A questo punto passiamo ad osservare la tabella degli **autovalori** e delle **percentuali di inerzia**:

Gli autovalori devono risultare inferiori a **1**, difatti ricordiamo dalla definizione di autovalore, come variabilità di variabili che devono essere standardizzate al fine di consentirne il confronto; dunque, da ciò deriva che l'inerzia/varianza di queste nuove variabili deve essere minore di **1**.

Consideriamo, adesso, la **tabella dei profili**, nel caso delle righe.

Nell'**AC** si pone particolare attenzione ai profili riga e ai profili colonna, inoltre geometricamente ogni profilo, riga o colonna, viene considerato come vettore in uno spazio multidimensionale.

A dispetto dell'**ACP**, in cui ogni punto ha un peso pari a $\frac{1}{n}$, nell'**AC** ogni punto ha una massa pari al rapporto tra il rispettivo marginale, di riga o di colonna, e il totale della tabella.

La caratteristica delle tabelle dei profili è quella di avere, come somma delle componenti, un valore costante (pari a **1** oppure a **100**); difatti, ciò ha come conseguenza la perdita di una dimensione del corrispondente spazio di riferimento.

Il profilo di riga, osservando la tabella dei profili delle righe, si calcola dividendo il valore f_{ij} per il marginale di riga $f_{i.}$.
L'ultima riga rappresenta il profilo riga medio dal quale traiamo alcune informazioni.

Una volta osservata la tabella dei profili riga, consideriamo le coordinate principali delle righe, in cui come sappiamo perdiamo una dimensione; inoltre, ricordiamo che nell'**AC** si ha un insieme di punti pesanti e che le coordinate si ottengono moltiplicando la matrice $\boldsymbol{X}$ per $\boldsymbol{M}$ per $\boldsymbol{U}$

$$\boldsymbol{Coordinate\ AFC = XMU}$$

$\boldsymbol{M}$ è una matrice che presenta, sulla diagonale, dei pesi, i quali risultano i reciproci delle frequenze relative.
Il prodotto $\boldsymbol{M}$ per $\boldsymbol{U}$ significa che la prima riga modifica il valore di $\boldsymbol{U}$, dunque la quantità $\boldsymbol{M}$ per $\boldsymbol{U}$ prende il nome di fattore.
Proprietà fondamentale, in questo caso, è che il prodotto tra $\boldsymbol{U'\,M}$ e $\boldsymbol{U}$ risulti pari a $\mathbf{1}$

$$M : \begin{bmatrix} 1/f_1 & & \\ & \ddots & \\ & & 1/f_r \end{bmatrix} \begin{bmatrix} u_1 \\ \\ u_r \end{bmatrix}$$

La formula per il calcolo delle coordinate può essere anche scritta come:

$$\boldsymbol{Coordinate\ AFC = X\sum^{-1} U}$$

Dove $\boldsymbol{\Sigma^{-1}}$ è la $\boldsymbol{M}$ della formula precedente, vista stavolta come prodotto di deviazioni standard.
Si nota che $\boldsymbol{\Sigma}$ è elevato alla $\boldsymbol{-1}$, ciò significa che è diviso per la deviazione standard, facendo, dunque, una standardizzazione.

Osserviamo, poi, la **tabella delle coordinate principali** delle righe.

In seguito, possiamo riferirci alla tabella dei coseni al quadrato per le righe, occorre però definire questi ultimi.

Il **coseno al quadrato** è una misura della bontà della rappresentazione del punto nel sottospazio di riferimento, difatti un punto sarà tanto meglio rappresentato, sull'asse, quanto più il valore del coseno al quadrato si avvicina ad **1**; quindi, tanto più piccolo sarà l'angolo formato dai 2 vettori tanto migliore risulterà la rappresentazione.
Dalla formula,

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

il quadrato del coseno dell'angolo formato dal *vettore proiezione* e dal vettore nello spazio originario, può essere calcolato rapportando le norme corrispondenti; infatti, tale formula altro non è che è il rapporto tra due lunghezze al quadrato, più nello specifico il cateto e l'ipotenusa della proiezione del punto sul nuovo sistema di assi.

A questo punto osserviamo la tabella dei **coseni quadrati** delle righe.

Se i valori dei coseni al quadrato risultano più alti per i primi due assi ($F1$ e $F2$), ciò sta a significare che questi ultimi spiegano la maggior parte della distribuzione, ben rappresentando il fenomeno.

Avendo osservato i profili delle righe non ci resta che analizzare i profili delle colonne, ricordiamo a tal proposito che il profilo colonna si calcola dividendo il valore f_{ij} per il marginale di colonna $f_{.j}$.
Successivamente osserviamo la **tabella dei profili** delle colonne.

Riguardo la tabella dei profili colonna analizziamo se sia possibile o meno applicare la proprietà dell'**equivalenza distributiva**, secondo cui sia possibile unificare due profili riga uguali o proporzionali senza che tale trasformazione aggiunga informazioni, ossia non mutando i risultati.

Nella tabella analizzata utilizzare l'equivalenza distributiva comporterebbe accorpare tutti i valori della matrice, essendo tra loro molto simili; dunque, non risulta possibile attuare la suddetta equivalenza distributiva.

Ricordiamo che nell'**AC** abbiamo un insieme di punti pesanti, quindi l'autovalore risulta pari a:

$$\lambda = \Sigma \varphi^2{}_i f_i = \frac{\varphi^2{}_i f_i}{\lambda}$$

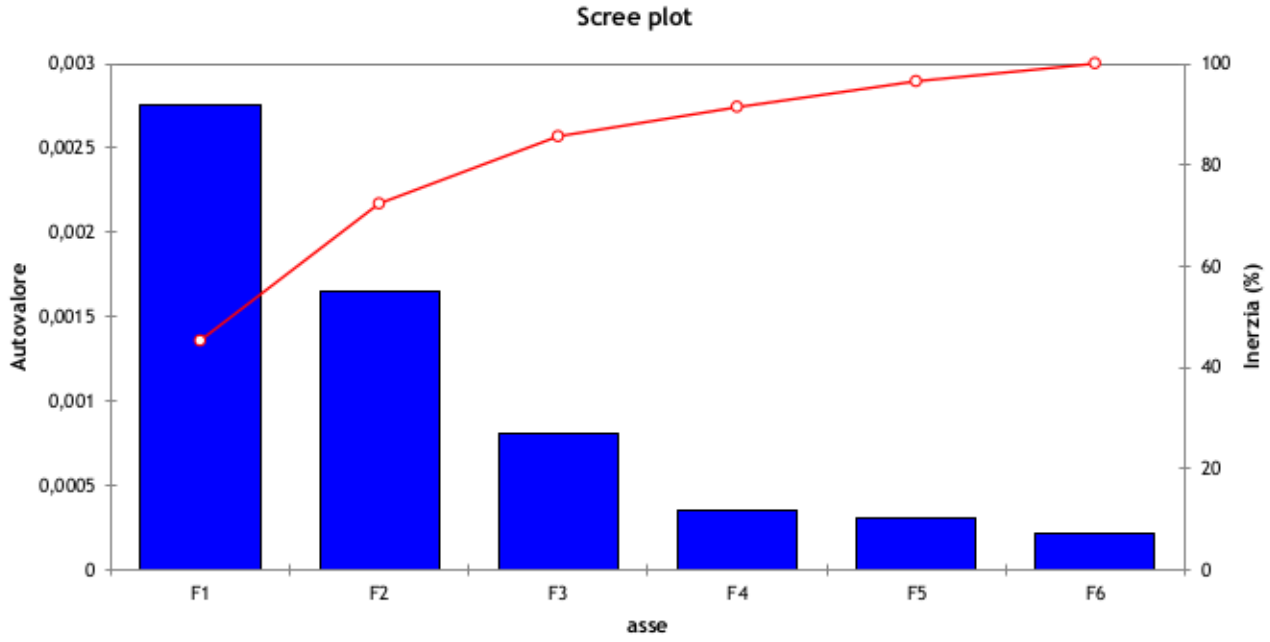
dove f_i rappresenta il peso. Possiamo affermare, inoltre, che più grande è la coordinata pesata, maggiore è il contributo; difatti, a parità di coordinate il contributo è maggiore qualora il peso risulti maggiore.

Allora, possiamo osservare la tabella dei contributi delle colonne.

Consideriamo i coseni al quadrato, stavolta per le colonne, considerando la tabella dei **coseni quadrati**.

Se i valori dei coseni al quadrato risultano più alti per i primi due assi ($F1$ e $F2$), ciò sta a significare che questi ultimi spiegano la maggior parte della distribuzione, ben rappresentando il fenomeno.

A questo punto, per suffragare quanto osservato nella tabella, riguardo la rappresentatività degli assi nello spiegare le variabili, possiamo osservare lo **Scree plot**

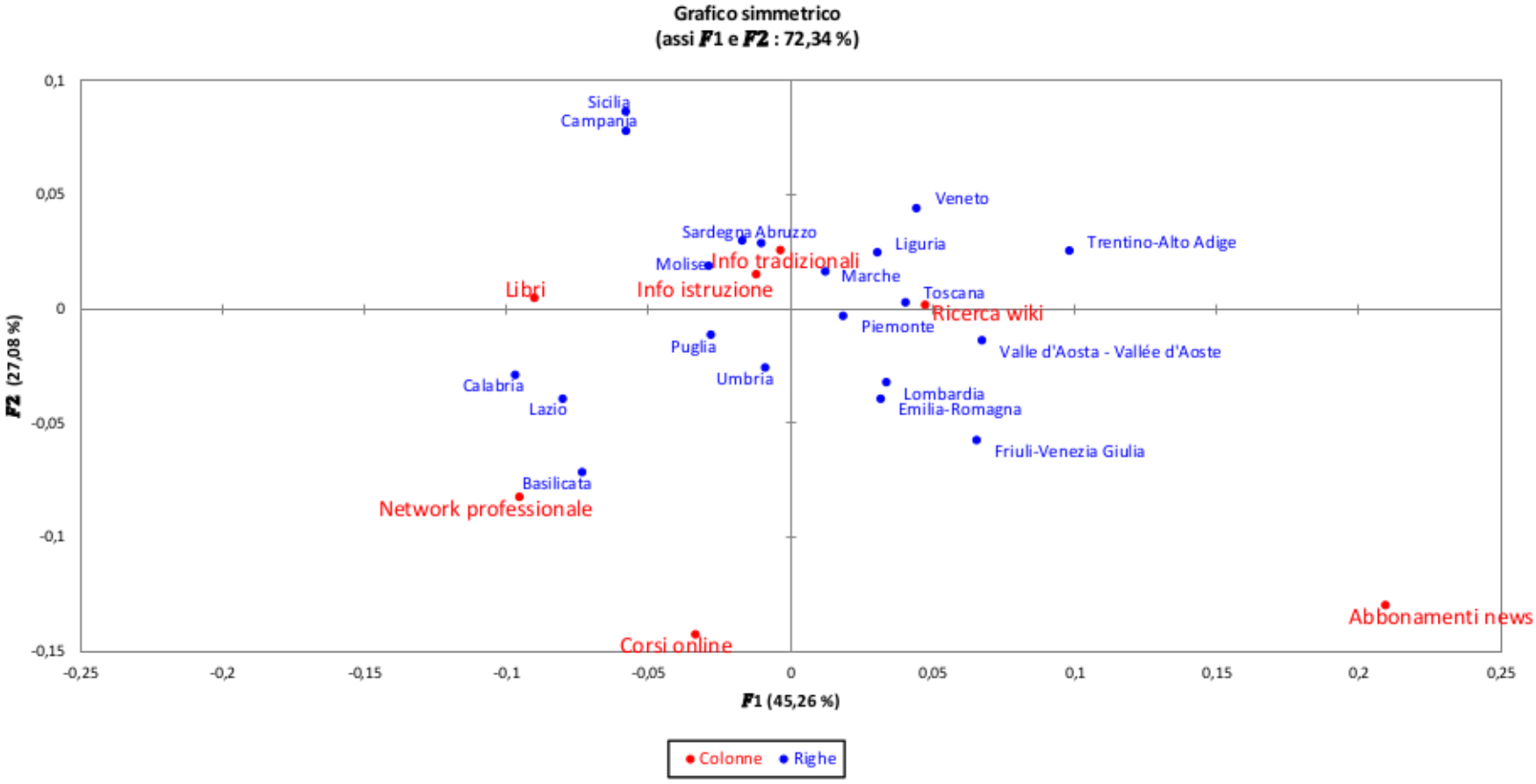


Osservando lo **Scree plot** abbiamo la conferma di quanto avevamo intuito vedendo la tabella dei coseni al quadrato, ossia che gli assi ***F*1** e ***F*2** ben rappresentano il fenomeno, perché dopo di questi si nota uno *scalino*; tuttavia, non si può osservare un salto considerevole, essendoci già una sensibile differenza tra l'asse **F1** e l'asse **F2**.

I risultati dell'**AC** possono essere rappresentati su grafici piani ottenuti utilizzando coppie dei fattori individuati, tuttavia differentemente dall'**Analisi in Componenti Principali**, l'**AC** recupera una simmetria complessiva come conseguenza di una ideale sovrapposizione dei risultati di due analisi asimmetriche.

La tabella dei profili riga ci consente di posizionare le regioni italiane all'interno della nube definita dai profili colonna. Questa proprietà è detta **proprietà baricentrica** dell'Analisi delle Corrispondenze, calcolando la posizione di ogni punto-profilo come media ponderata dei punti-vertice definiti dalla modalità dell'altro carattere.

Osserviamo, dunque, il **grafico simmetrico**



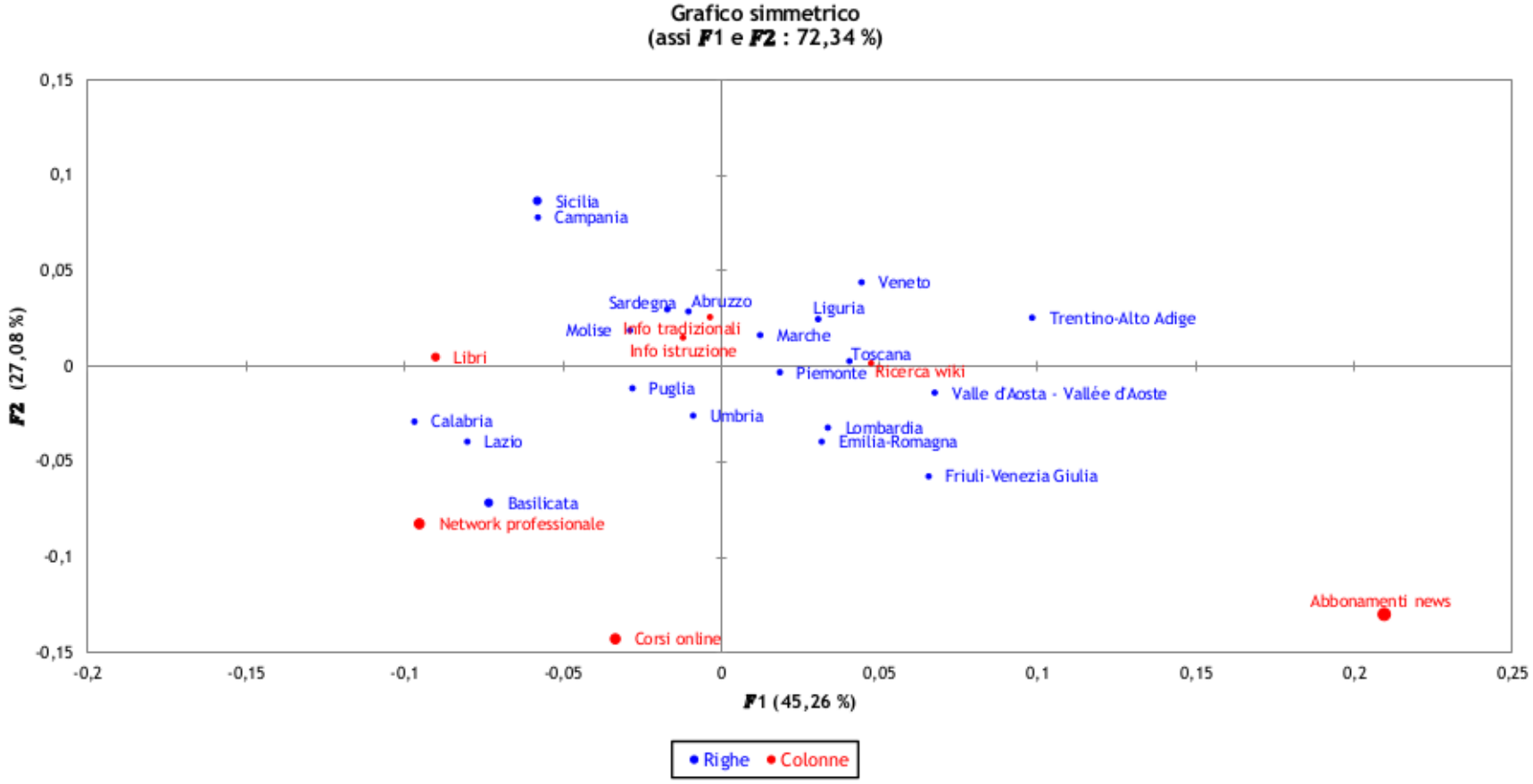
La lettura del grafico è resa possibile dal fatto che la percentuale di inerzia, spiegata dai primi due fattori, risulta pari al 72,34%; invece, se la percentuale fosse stata molto elevata (ad esempio pari al 95%), la lettura dei piani sarebbe stata meno immediata, essendoci un numero elevato di modalità delle variabili, in quanto molti punti risultano spesso sovrapposti.
L'**AC** consente allora l'ideale sovrapposizione dei piani fattoriali generati dalle analisi nei due spazi e, quindi, la rappresentazione simultanea dei profili; difatti, tale rappresentazione viene definita $\boldsymbol{\beta - baricentrica}$, al fine di sottolineare la diversità dalla situazione ideale (impossibile) in cui ogni riga risulta baricentro delle colonne ogni colonna risulta baricentro delle righe.

La rappresentazione simultanea delle modalità dei due caratteri permette di valutarne le relazioni.
Una volta individuati gli assi principali su cui proiettare i punti, è necessario determinare dei coefficienti che consentano una corretta interpretazione dei grafici e che tengano, quindi, conto da un lato delle prossimità tra punti e piani principali e dall'altro del ruolo giocato da ciascun punto nella determinazione degli assi; poiché, le variabili trattate sono di tipo qualitativo, alcune delle qualità note dall'**ACP** non potranno essere utilizzare (ad esempio le correlazioni variabili-fattori), mentre per altre sarà possibile estenderne per analogia i principi metodologici.

Si definisce **contributo assoluto** del punto $\boldsymbol{i}$ (o del punto $\boldsymbol{j}$) all'$\alpha - esimo$ asse la percentuale di variabilità dell'asse spiegata dal punto $\boldsymbol{i}$ (o dal punto $\boldsymbol{j}$).
Una misura della qualità della rappresentazione dei punti sugli assi (talvolta detta **contributo relativo**) è, invece, data dal quadrato del coseno dell'angolo formato dal vettore proiezione del punto $\boldsymbol{i}$ (o del punto $\boldsymbol{j}$) e il vettore relativo al punto $\boldsymbol{i}$ (o al punto $\boldsymbol{j}$) nel proprio spazio originario.

Il coseno al quadrato è un indice della bontà della rappresentazione del punto nel sottospazio di riferimento, difatti un punto sarà tanto meglio rappresentato sull'asse quanto più il valore del coseno al quadrato si avvicina a **1**.

Osserviamo di seguito il **grafico simmetrico** con la visualizzazione dei **contributi** delle variabili



Dal grafico simmetrico possiamo osservare che le regioni italiane a destra del baricentro hanno una media superiore, mentre quelle a sinistra del baricentro hanno una media inferiore.
Le regioni presenti nel quadrante in basso a sinistra hanno una media negativa rispetto ad entrambi gli assi.