L'Analisi dei Dati Testuali

dalle "parole" ai "fatti"...

Michelangelo Misuraca – università della calabria
dipartimento di economia e statistica

Grazie ai progressi dell'informatica e alla maggiore disponibilità di dati, negli anni '60 la Statistica incomincia un processo di rinnovamento che modifica profondamente i rapporti fra teoria (*modello*) ed osservazione (*dati*): nasce così l'**analisi dei dati**

Uno dei nuovi approcci è quello sviluppato nell'ambito della cosiddetta *scuola francese di analisi multidimensionale dei dati* (AMD) ed è legato al nome di J.P. Benzécri

Statistica classica:
- Obiettivo confermativo (modello-dati)
  campione casuale – inferenza statistico

Analisi dei Dati:
- Obiettivo esplorativo (dati-modello)
  EDA, J. Tukey
- Descrittivo (dati-struttura)
  AMD, J.P. Benzécri

L'AMD si propone di interpretare e visualizzare la struttura di fenomeni complessi mediante il trattamento di numerose **variabili** e **osservazioni**

Fornisce, infatti:
- la **visualizzazione di associazioni** anche complesse
- la definizione di **fattori multidimensionali**
- la costruzione di **tipologie** di osservazioni
- il disegno di **mappe**

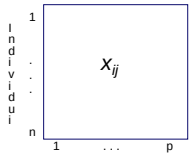
In sintesi con l'AMD si hanno due obiettivi principali. Dato un certo fenomeno si vuole:

Ricercare una struttura latente

Visualizzare questa struttura

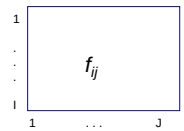
LA STRUTTURA DEI DATI

- **Osservazioni per variabili**

$\mathbf{X} =$ 

x_{ij} è il valore assunto dall'i-esimo individuo per la j-esima variabile quantitativa

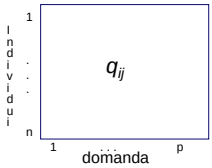
- **Tabelle di contingenza**

$\mathbf{F} =$ 

f_{ij} *frequenza congiunta*

$f_{i.}$, $f_{.j}$ *frequenze marginali*

- **Dati di questionario**

$\mathbf{Q} =$ 

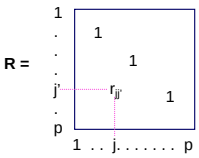
q_{ij} è la modalità di risposta fornita dall'i-esimo individuo alla j-esima domanda del questionario

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

LE MATRICI DI BASE DELL'AMD

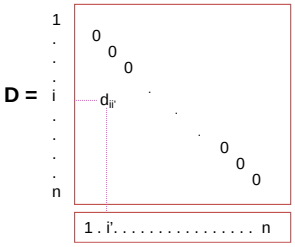
- **Matrici di correlazione (o di associazione)**

  Es. coefficienti di correlazione di Bravais-Pearson, covarianze, indici di cograduazione di Spearman



Es. coefficiente di correlazione lineare

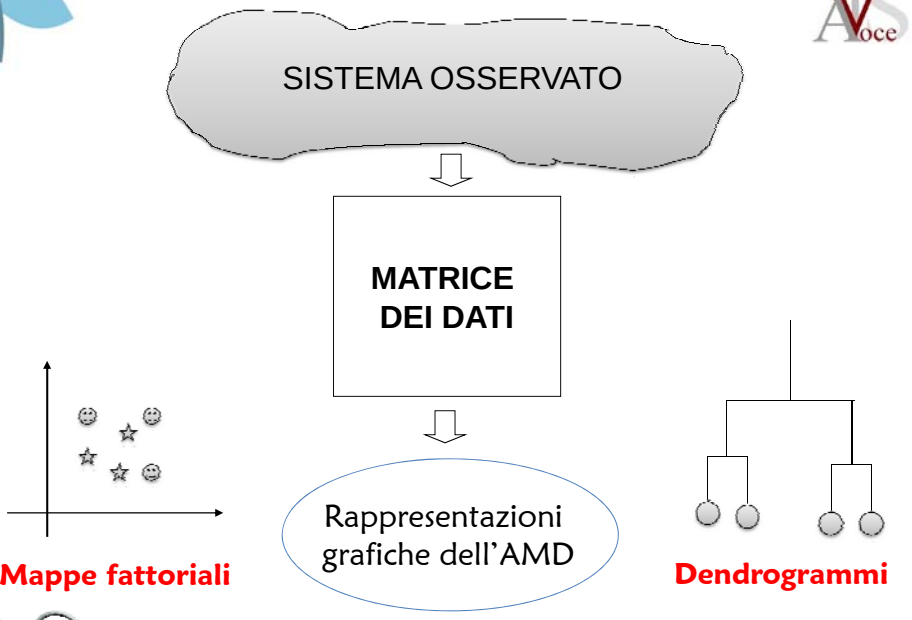
- **Matrici di distanze (o di similarità)**



Definizione di DISTANZA d:

- d(i,i) = 0
- d(i,i') ≥ 0
- d(i,i') = d(i',i)
- d(i,i') ≤ d(i,h)+d(i',h)

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali



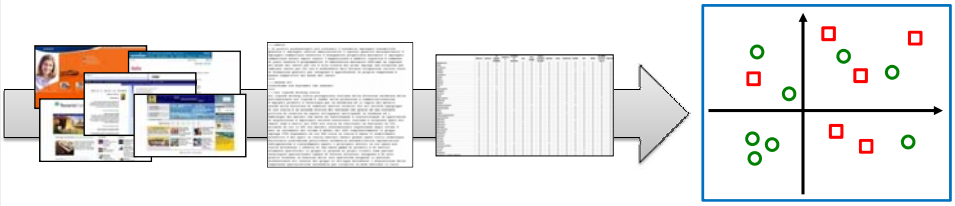
L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

ANALISI DELLE CORRISPONDENZE

Obiettivo:
Studiare la struttura dell'associazione tra due o più variabili qualitative

L'AC permette di decomporre una tabella a due o più entrate in una serie di **fattori**, ciascuno dei quali rappresenta un aspetto "latente" dell' 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 dalla lettura della tabella



La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

ANALISI DELLE CORRISPONDENZE LESSICALI

Per quanto detto, è possibile analizzare una tabella lessicale del tipo DxP (documenti x parole) attraverso l'analisi delle corrispondenze:

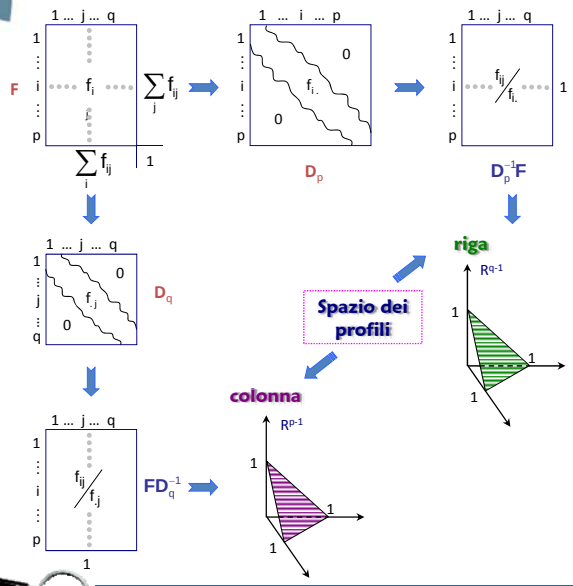
1) ricercare e visualizzare strutture linguistiche latenti per evidenziare la presenza di concetti o temi prevalenti

2) ricercare e visualizzare similarità tra documenti (in termini di vocabolario condiviso) per evidenziare la presenza di gruppi

Da un punto di vista concettuale sarebbe più corretto analizzare tabelle lessicali aggregate, ma è pratica comune quella di utilizzare nell'analisi anche le tabelle lessicali

L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

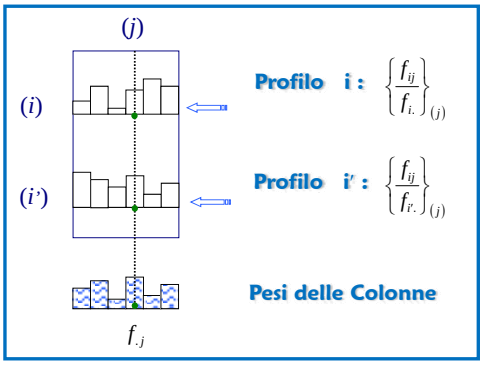
STRUTTURA DEI DATI



$$F = U \Lambda V^T$$
$$U^T D_p^{-1} U = V^T D_n^{-1} V = I$$

Obiettivo è trovare il miglior sottospazio di rappresentazione: vogliamo cioè rappresentare in due dimensioni la nube dei punti conservando però quanta più informazione possibile

MISURARE LE DISTANZE TRA DOCUMENTI O PAROLE



$$d^2_{ii'} = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Pesi di *j*

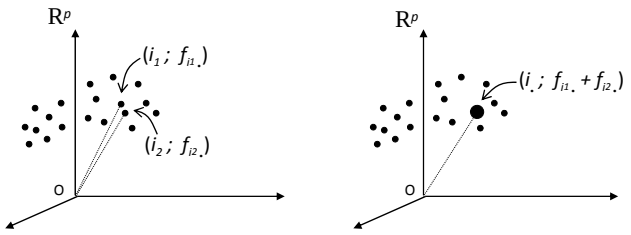
Scarto in colonne *j* tra i profili

$$d^2_{ii'} = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i}} - \frac{f_{i'j}}{f_{i'}} \right)^2$$

$$d^2_{jj'} = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

La metrica del Chi-quadro è costruita in modo da considerare una sorta di effetto "normalizzante" sull'importanza dei diversi punti

PRINCIPIO DELL'EQUIVALENZA DISTRIBUTIVA



Se i_1 e i_2 sono dei profili identici, i punti i_1 (pesi $f_{i1.}$) e i_2 (pesi $f_{i2.}$) sono confusi in R^p

Sia i_0 il punto comune, assegnato ai pesi $(f_{i1.} + f_{i2.})$ allora:

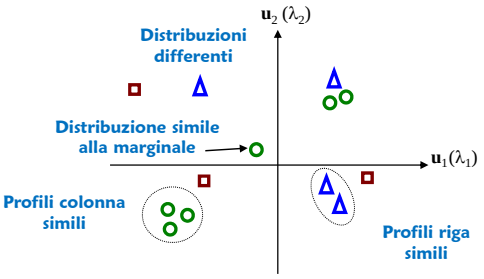
$\forall\ i$ e i' in R^p, $d^2(i,i')$ è invariato

$\forall\ j$ e j' in R^q, $d^2(j,j')$ è invariato

La Statistica come strumento di analisi nelle scienze umanistiche e comportamentali

LETTURA DELLA MAPPA

Profili riga
Profili colonna
Punti supplementari



1. la dispersione dei punti intorno all'origine mostra la forza dell'associazione
2. Se due parole sono vicine allora sono utilizzate in modo simile
3. Se due documenti o due modalità della variabile di classificazione sono vicine allora hanno un vocabolario simile
4. Non si può valutare la prossimità tra forme e documenti ma la posizione relativa di un punto «forma» rispetto alla nube dei documenti (e viceversa)
5. La dimensione delle coordinate suggerisce l'importanza di un punto rispetto all'asse

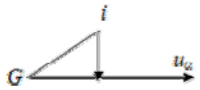
L'Analisi dei dati testuali: uno strumento per leggere tra le righe – Soriano nel Cimino 5-9 ott 2009

CONTRIBUTI ASSOLUTI E RELATIVI

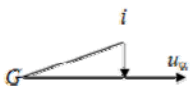
L'**inerzia totale** di una tabella misura la **disomogeneità** dei profili riga e dei profili colonna. Ogni riga e ogni colonna contribuiscono in relazione al loro allontanarsi dalla situazione di *indipendenza*, espressa dai marginali

I **contributi assoluti** ai singoli assi esprimono l'importanza delle modalità nei confronti di un fattore: si utilizzano per interpretarli più facilmente

I **contributi relativi** (o coseni quadrati) esprimono invece quanto un punto è deformato dalla proiezione sull'asse fattoriale: misurano quindi la qualità della rappresentazione e variano fra 0 e 1



i è mal rappresentato



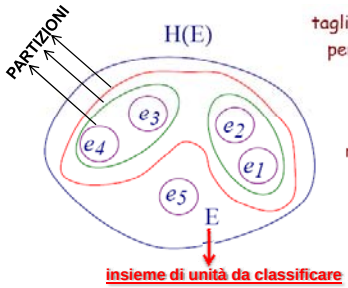
i è ben rappresentato

CATEGORIZZAZIONE DEI DOCUMENTI EX POST

Al termine dell'analisi fattoriale è possibile sulla base dei risultati ottenuti cercare di classificare automaticamente i documenti

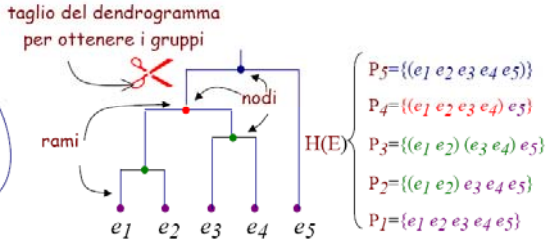
E' possibile ad esempio ricorrere ad una tecnica nota come **Cluster Analysis**

GERARCHIA



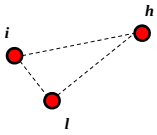
insieme di unità da classificare

DENDROGRAMMA



DISSIMILARITA', METRICA E ULTRAMETRICA

Per classificare un insieme di unità è necessario definire come raggruppare/separare chi è simile/dissimile oppure vicino/lontano



a) $d_{ii'} = \varnothing \Leftrightarrow e_i = e_{i'}$ (separabilità)

b) $d_{ii'} = d_{i'i}$ (simmetria)

c) $d_{ih} \leq d_{il} + d_{lh}$ (diseguaglianza triangolare)

d) $d_{ih} \leq \mathrm{Max}\{d_{il}, d_{lh}\}$ (condizione di Krassner)

d { INDICE DI DISSIMILARITÀ se verifica le condizioni a e b
DISTANZA O METRICA " " a, b e c
DISTANZA ULTRAMETRICA " " a, b e d

	e1	e2	e3	e4
e1	0	3	2	4
e2	3	0	6	1
e3	2	6	0	1
e4	4	1	1	0

indici di dissimilarità

	e1	e2	e3	e4
e1	0	7	6	7
e2	7	0	5	2
e3	6	5	0	6
e4	7	2	6	0

distanze

$d_{23} > d_{21} + d_{13}$

	e1	e2	e3	e4
e1	0	4	4	1
e2	4	0	3	4
e3	4	3	0	4
e4	1	4	4	0

ultrametriche

$d_{21} > \max d_{23}, d_{13}$

A seconda della natura dei dati è possibile definire diversi tipi di distanze

UNA POSSIBILE STRATEGIA INTEGRATA

Per poter analizzare e classificare automaticamente una collezione di documenti è necessario partire da una tabella (*documenti x forme*):

STEP 0 – pretrattamento del corpus e costruzione della tabella lessicale

STEP 1 – Analisi delle Corrispondenze Lessicali

STEP 2 – Cluster Analisys sui fattori ottenuti dalla ACL

ATTENZIONE!!!
La classificazione ottenuta considera le diverse caratteristiche (in questo caso le forme utilizzate) in termini di UNIONE (OR) e non in termini di INTERSEZIONE (AND)

In fase di interpretazione sarebbe necessario ricorrere ad altri strumenti (*marcaggio simbolico* o *regole di associazione*)