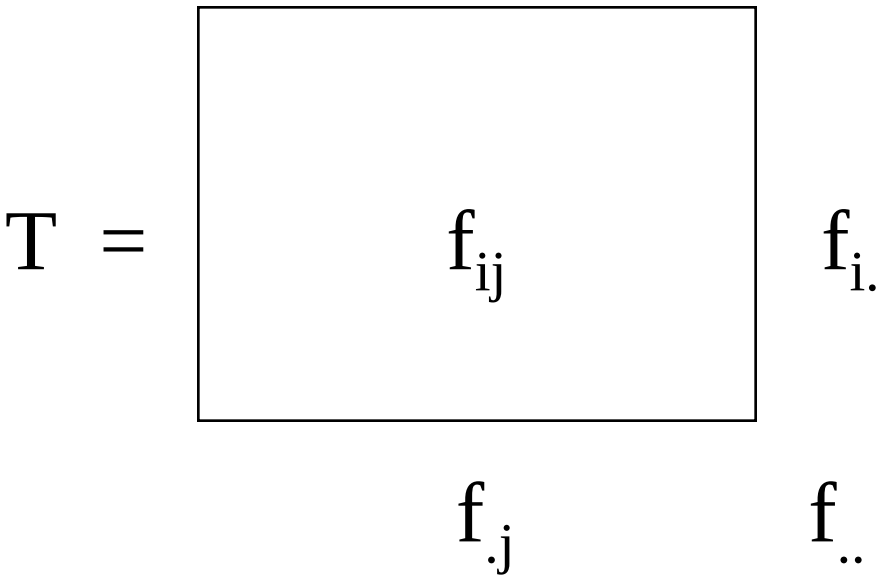
Quale *tabella testuale*?



Quali *unità statistiche*?

Quali sono le unità della statistica testuale?

- Una volta definiti i caratteri *delimitatori* (spazi, virgole, ecc.) si definisce **occorrenza** una *sequenza di caratteri all'interno di due delimitatori*
- Sequenze identiche costituiscono le occorrenze di una **forma**
- L'insieme delle forme in un *corpus* ne costituisce il **vocabolario**
- Esistono vari tipi di forme: *grafiche*, *lemmatizzate* (in italiano, tutti i verbi all'infinito, tutti i sostantivi al singolare, tutti gli aggettivi al maschile singolare), *testuali*

Ancora sulle unità dell'analisi dei dati testuali

- analisi dei dati testuali possono essere condotte non su "parole", ma su sequenze di occorrenze
- si tratta dei cosiddetti **segmenti ripetuti**, ossia sequenze di occorrenze non interrotte da *separatori forti* (cioè, di frase: . ? ! ;), che si presentano identiche nel *corpus* con una frequenza superiore ad una soglia prefissata

  es. livello di vita
- esistono, ancora, i **quasi-segmenti ripetuti**,

  es. per i figli; per i miei figli; per i nostri figli; per i suoi figli; per i propri figli

La costruzione della tabella lessicale - 1

Occorre contare le unità che si vogliono analizzare, considerandole realizzazioni di una forma generale

L'operazione che consente di decomporre il testo in unità elementari è detta **segmentazione**

Il successivo raggruppamento delle unità considerate identiche (*realizzazioni di una stessa forma*) è detta fase della **identificazione**

Si noti che questi due procedimenti sono regolati da **norme** dettate dallo specifico campo di applicazione e dalla stessa definizione di forma prescelta

La costruzione della tabella lessicale - 2

- A tutte le occorrenze di una delle V forme presenti nel *corpus* viene associato un numero
- Il numero di occorrenze in una porzione di testo considerata, detta sub-testo, o parte, è considerata la sua lunghezza. Si tratta, ad esempio, della risposta ad una domanda aperta di un questionario, fornita dagli N intervistati
- Indicando con t_j la lunghezza della j-esima parte (es.risposta), poiché ogni occorrenza appartiene ad una ed una sola parte, la lunghezza del testo è pari a:

$$T = \sum_{j=1}^{N} t_{ij}$$

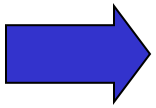
La tabella lessicale

- Una **tabella lessicale** **T**, di dimensioni (N,V) è una particolare tabella di contingenza il cui generico elemento t_{ij} è il numero di occorrenze della forma j all'interno del sub-testo i
- E' così possibile realizzare un'analisi delle corrispondenze della tabella lessicale
- Problema: i sub-testi hanno lunghezza variabile e, come tabella di contingenza, **T** ha dimensioni molto grandi e un numero elevato di 0

T è povera di informazione

La tabella lessicale aggregata

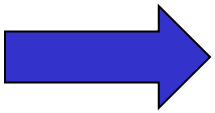
E' allora interessante analizzare i sub-testi aggregati rispetto ad una (o più) caratteristiche ritenute rilevanti da un punto di vista testuale. Nel caso di analisi di dati di questionario, si considerano, in genere, gruppi di intervistati definiti rispetto alla loro risposta ad una (o ad un mix) di domande chiuse



Costruire una matrice ***C*** che incrocia la partizione dei sub-testi e le forme utilizzate all'interno dei gruppi di sub-testi considerati
E' questa la matrice su cui si effettuerà AC

Costruzione di **C**

Per ottenere la nuova matrice **C** (*V*,*P*) partizione-vocabolario si introduce una matrice indicatrice **Z** (*N*,*P*), con in riga i sub-testi e in colonna le modalità della variabile che induce la partizione. **Z** presenta in ogni riga tutti 0, tranne un 1 in corrispondenza del gruppo cui appartiene il sub-testo



$$\underset{(V,P)}{C} = \underset{(V,N)}{T'}\ \underset{(N,P)}{Z} \qquad \text{con } P \ll N$$

C ha *V* righe (*forme*) e *P* colonne (*gruppi*). L'elemento generico cj_k è il numero delle volte in cui la j-esima forma è stata usata nei sub-testi che presentano la *k*-esima modalità della variabile di partizione

La pubblicità dei vini

Messaggi: pubblicità di vini, dal 1992 al 1994, sul *Gambero rosso*
Obiettivo: relazione fra regioni e modelli culturali di riferimento

Regioni	Messaggi	Forme grafiche	F.g./M.
Piemonte + Val d'Aosta	16	553	35,6
Veneto(+1 Lombardia)	16	842	52,6
Friuli - Venezia Giulia	8	409	51,1
Trentino-Alto Adige	6	246	41,0
Emilia - Romagna	6	402	67,0
Toscana	20	1531	76,6
Marche	4	73	18,3
Abruzzo + Molise	2	40	20,0
Campania + Calabria	2	130	65,0
Sicilia	8	179	22,4
Sardegna	3	119	39,7
TOTALE	91	4524	49,7

Lemmatizzazione:

- Singolare per i nomi,
- Singolare maschile per gli aggettivi,
- Infinito per i verbi.

Criteri di eliminazione:

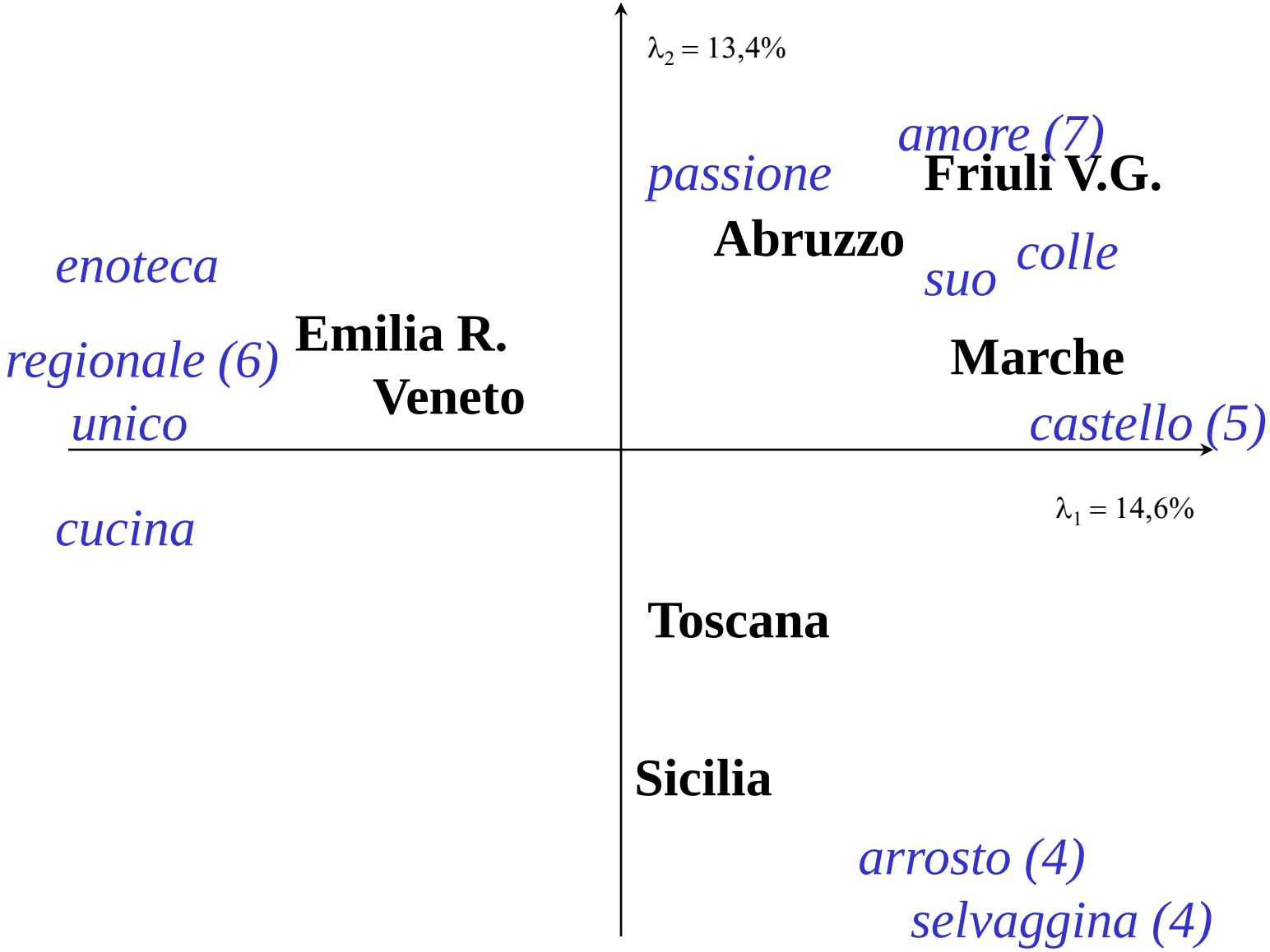
preposizioni, articoli (parole strumentali), tutte le parole usate meno di 4 volte.

Come si legge il piano fattoriale di una AC

Il grafico più utilizzato nell'AC, anche testuale, è il joint plot, in cui ciascun punto riga è quasi-baricentro delle modalità della variabile in colonna, e viceversa (formule di uransizione) .

Esistono alcune regole di lettura da rispettare:

- la dispersione dei punti intorno all'origine mostra la forza dell'associazione nella tabella di contingenza
- se due parole sono vicine sono utilizzate in maniera simile
- se due modalità della variabile di partizione sono vicine, allora vuol dire che utilizzano un vocabolario simile
- non si può leggere la prossimità di una parola ad una modalità, (o viceversa), ma valutare la sua posizione rispetto all'intera nube delle modalità (o parole)
- la grandezza delle coordinate suggerisce l'importanza di un punto rispetto all'asse corrispondente, da verificare con i contributi assoluti



Primo piano fattoriale dell'Analisi delle Corrispondenze

L'analisi non simmetrica

L'analisi **non simmetrica** delle corrispondenze invece di decomporre il ϕ^2, ossia una misura di associazione, decompone un indice di predicibilità, il τ_b di Goodman e Kruskal:

$$\tau_b = \frac{\sum_{k=1}^{P} \sum_{j=1}^{V} \left[\left(f_{jk} - f_{j.} f_{.k} \right)^2 / f_{.k} \right]}{1 - \sum_{j=1}^{V} f_{j.}^2}$$

con l'obiettivo di analizzare la **dipendenza** del vocabolario dalle caratteristiche dei sub-testi.

Analisi non simmetrica delle corrispondenze (ANSC)

- L'ANSC studia le *j* distribuzioni condizionate $f_{jk}/f_{.j}$ rispetto all'ipotesi di indipendenza $f_{j.}$
- In termini geometrici, questo si traduce nel calcolo delle distanze fra le modalità della variabile condizionante secondo la consueta metrica euclidea, invece di quella del χ^2 (mentre le distanze fra le modalità della variabile condizionata sono euclidee ponderate). Il vantaggio è di non attribuire peso eccessivo alle modalità rare.

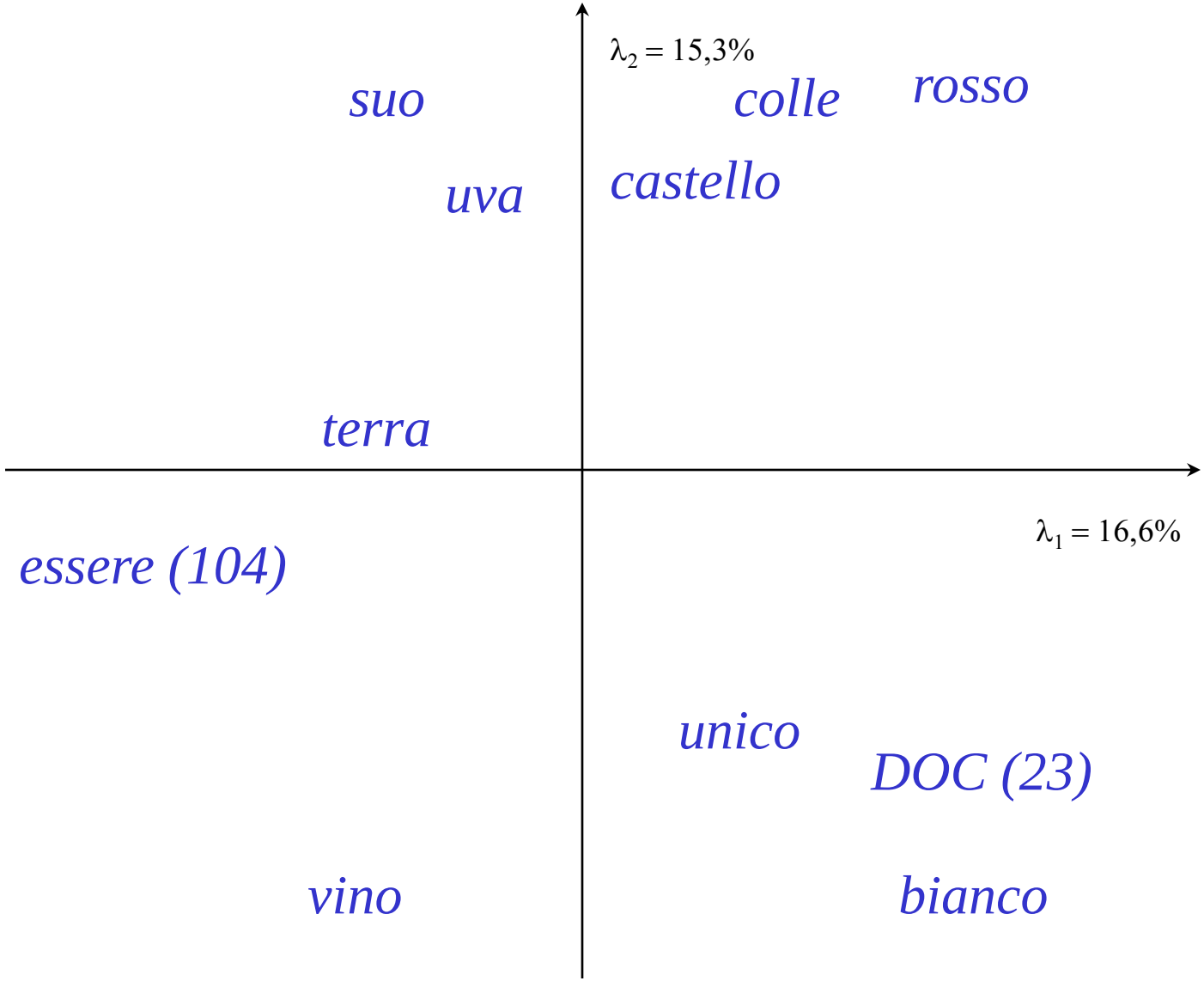
Questo e' il caso delle parole

Come si legge il piano fattoriale di una ANSC

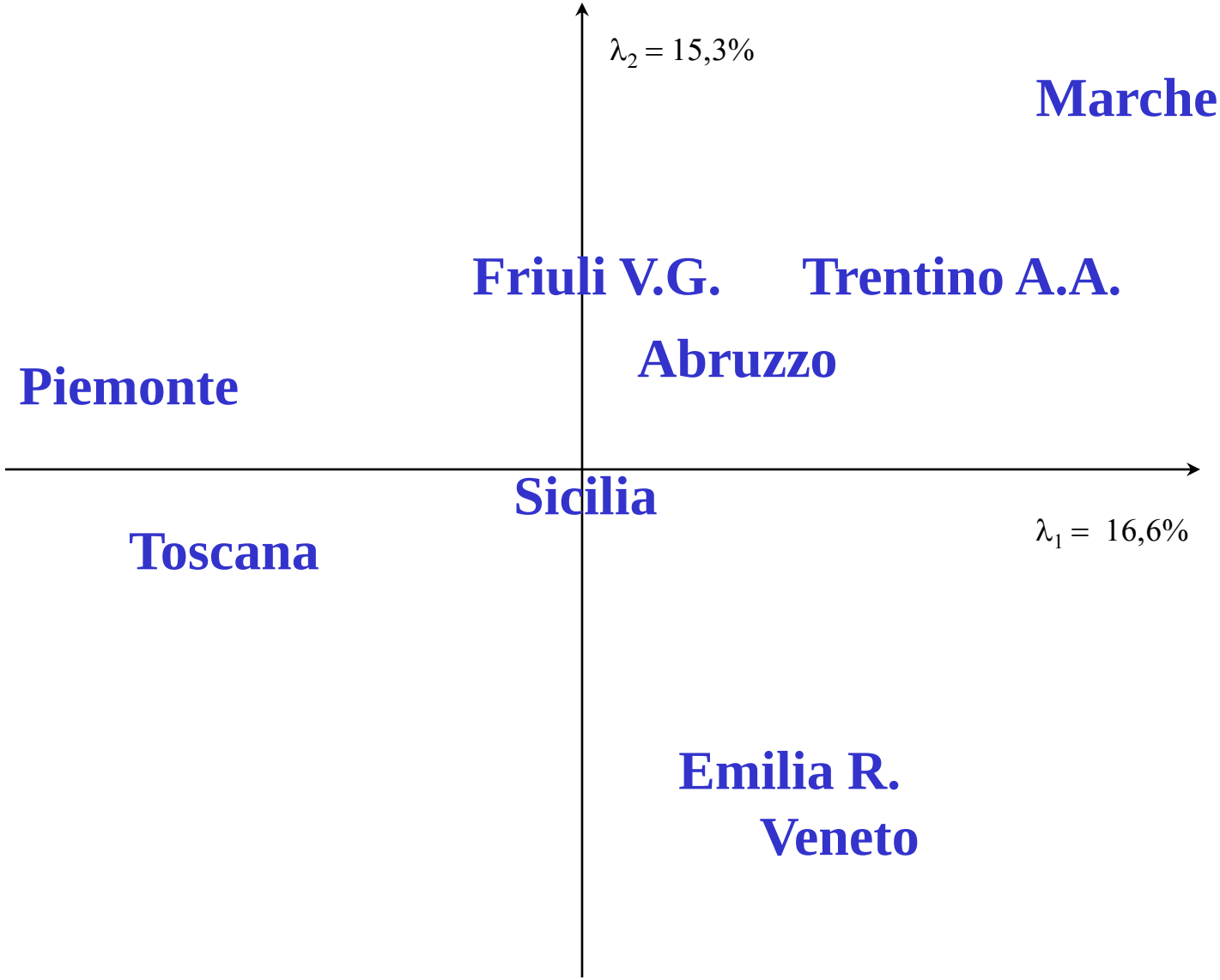
Nell'ANSC non è conveniente ricorrere al joint plot (le metriche nei due spazi sono molto differenti) anche se l'origine è comune e rappresenta, come per l'AC, l'ipotesi di indipendenza. Si ricorre a due rappresentazioni baricentriche separate.

Anche in questo caso esistono alcune regole di lettura:

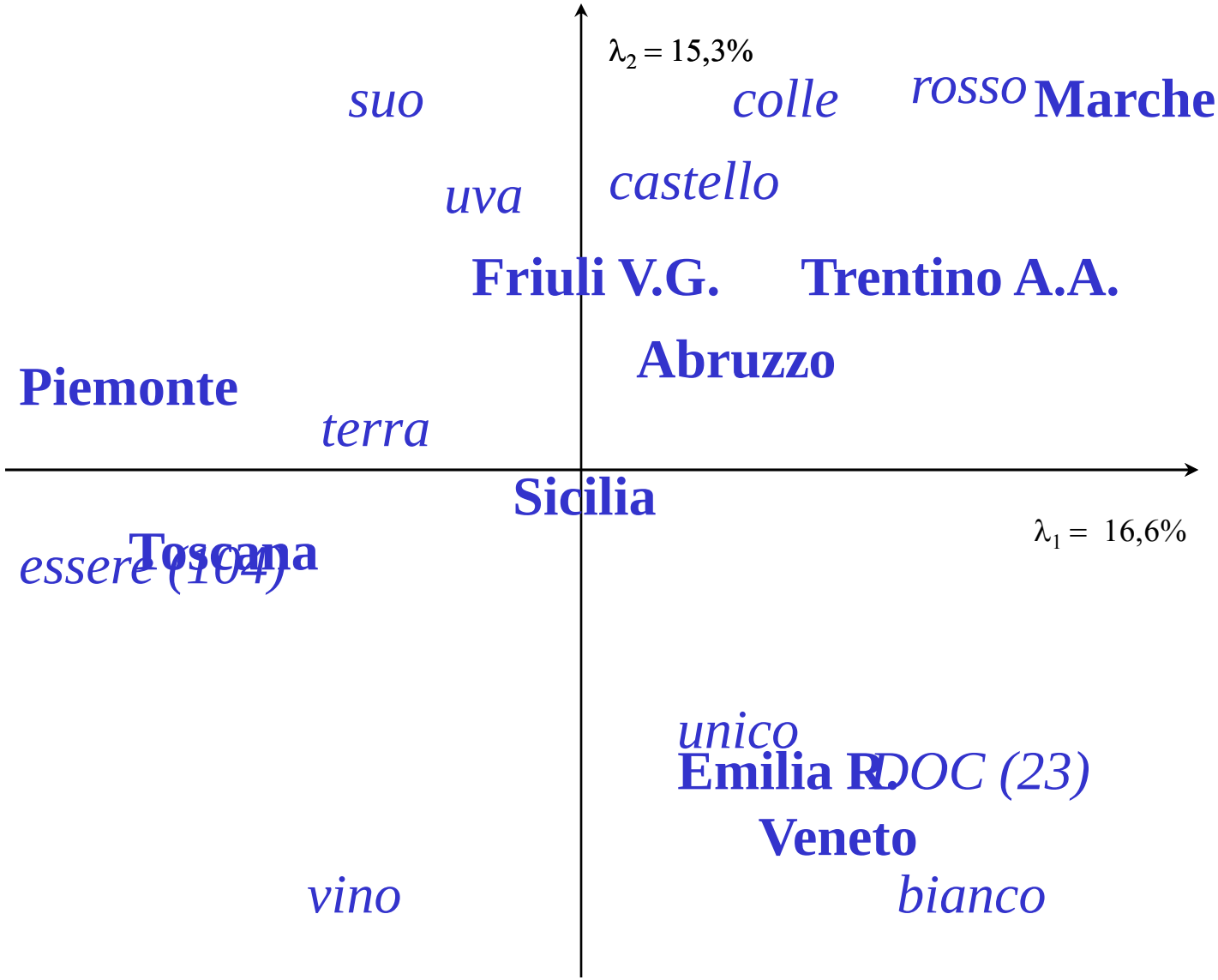
- la dispersione delle parole intorno all'origine mostra la forza della dipendenza del vocabolario dalla partizione
- due parole sono vicine se dipendono similmente dalla partizione
- due modalità della variabile di partizione sono vicine se influenzano in maniera simile l'uso delle parole
- una parola è tanto più lontana dall'origine quanto più è dipendente
- una modalità della variabile di partizione è tanto più lontana dall'origine quanto più influenza l'uso delle parole



Primo piano fattoriale dell'Analisi Non Simmetrica delle Corrispondenze



Primo piano fattoriale dell'Analisi Non Simmetrica delle Corrispondenze


$\lambda_2 = 15{,}3\%$
suo
colle
rosso
Marche
uva
castello
Friuli V.G.
Trentino A.A.
Abruzzo
Piemonte
terra
Sicilia
$\lambda_1 = 16{,}6\%$
essere (104)
Toscana
unico
Emilia R.
DOC (23)
Veneto
vino
bianco

Alcuni strumenti per l'analisi fattoriale di tabelle lessicali aggregate

- *Problemi di disambiguazione: i bootstrap convex hulls*
- *Introduzione di informazioni esterne sulle parole*
- *Introduzione della variabile tempo: un'analisi di matrici testuali a più vie*

Uno strumento computazionale per scegliere le unità testuali

per prendere decisioni in presenza di problemi di *disambiguazione* come, ad esempio, decidere:

- se le diverse **forme grafiche** di uno stesso morfema hanno una differente rappresentazione sui piani fattoriali (*problemi di lemmatizzazione, v. anche l'indice di iso-frequenza di Bolasco*)
- analogamente, se i **sinonimi** sono vicini sui piani fattoriali (*problemi di contestualizzazione*)

è possibile ricorrere ad uno **strumento computazionale**, utilizzato anche per la valutazione dei risultati di un'analisi delle corrispondenze.

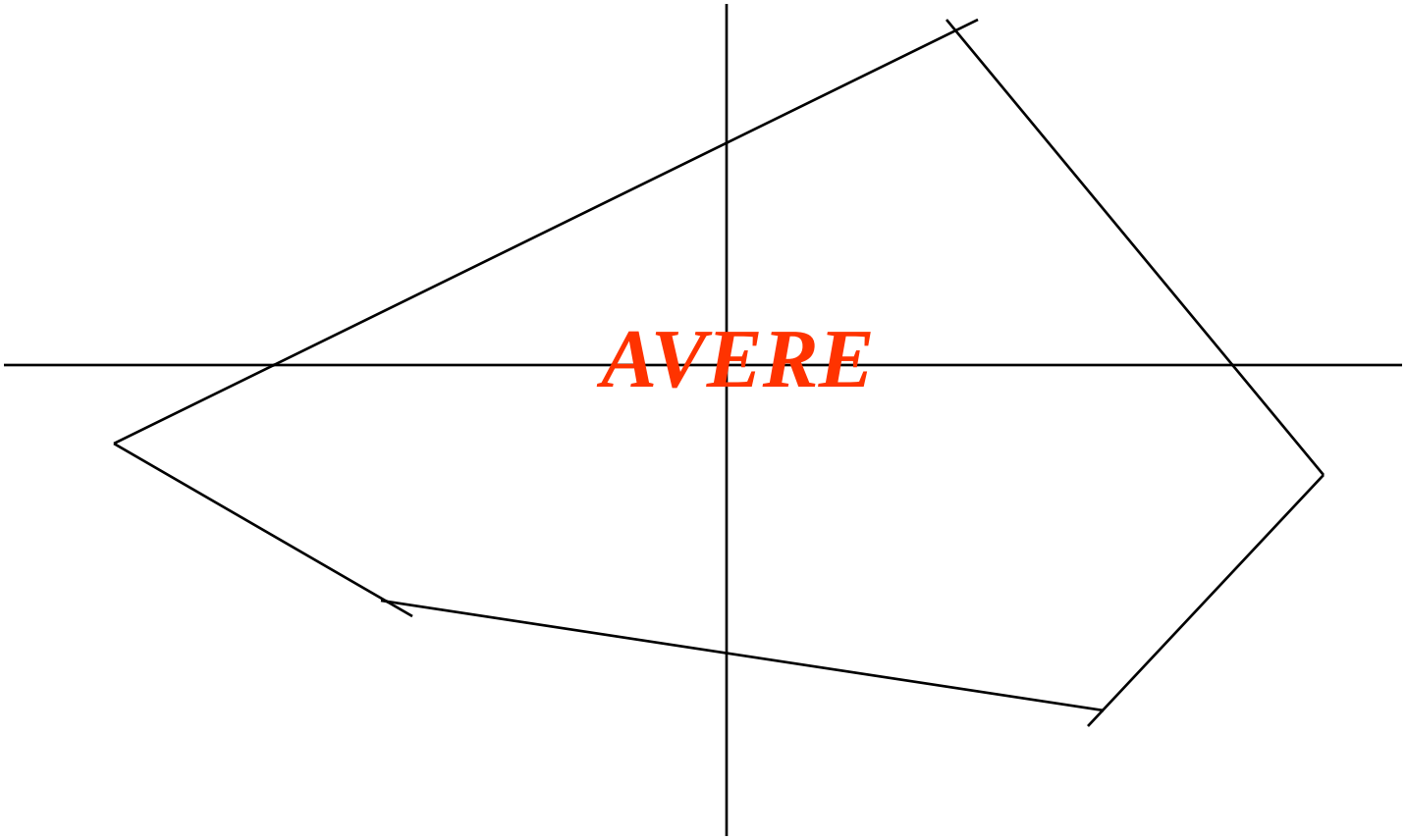
BOOTSTRAP CONVEX HULLS

- Si **ricampiona con reimmissione** all'interno della matrice di dati osservata
- Si costruisce un numero elevato di tabelle replicate
- Si proiettano in supplementare su una matrice scelta come riferimento (matrice originaria, oppure una matrice media o un compromesso) tutti i punti relativi alle forme considerate
- Si identifica la superficie in cui cade una quota fissata (es. 95%) di punti relativi alla stessa forma.
- Si uniscono i punti più esterni (**Convex Hulls**)
- Si vede se i convex hulls di forme differenti sono distinti oppure si sovrappongono.

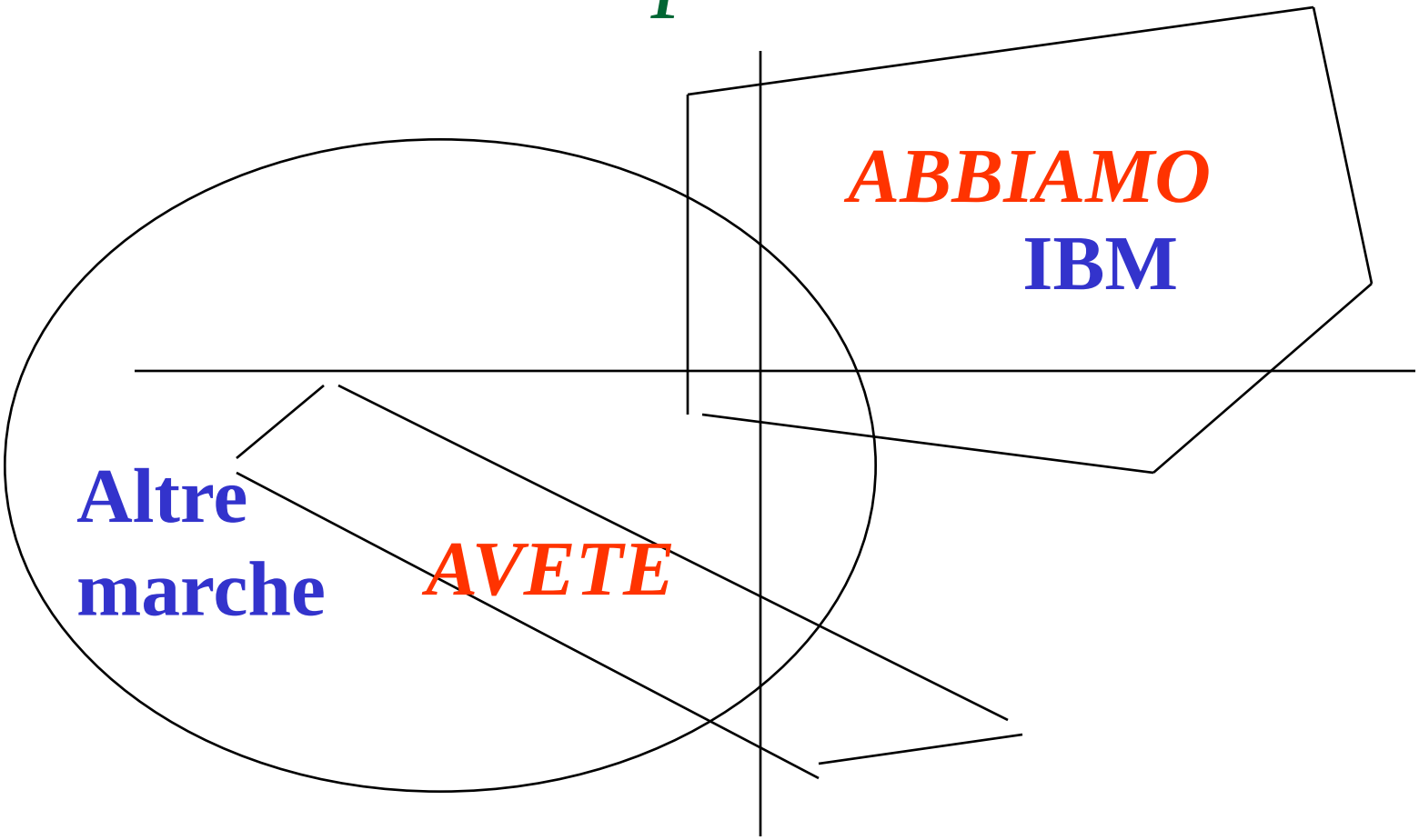
Le pubblicità dei computer

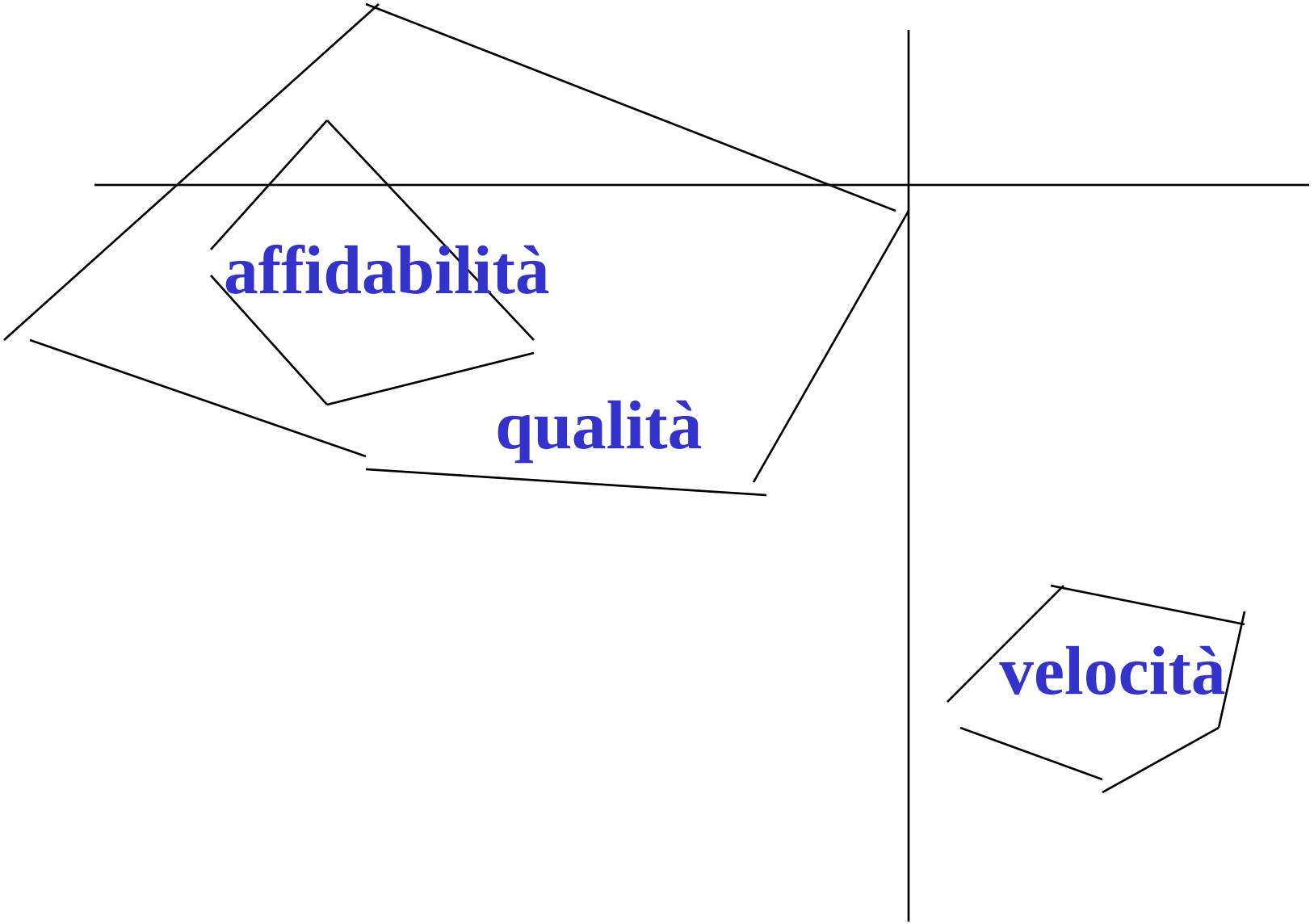
- Messaggi pubblicitari apparsi sui maggiori settimanali politico-economici italiani nel corso del 1995
- IBM *leader* : abbiamo
- Concorrenti: avete
- Forma lemmatizzata: avere

Strumenti per la disambiguazione



Bootstrap Convex hulls




affidabilità
qualità
velocità