Un'analisi dei dati testuali con informazioni esterne: le definizioni di "qualità"

Simona Balbi

Dipartimento di Matematica e Statistica, Università di Napoli "Federico II"

Giuseppe Giordano

Dipartimento di Istituzioni Economiche e Finanziarie, Università di Macerata

Abstract

Correspondence analysis is usually applied in order to identify and represent the association structure in a lexical table, **T**. **T** is a peculiar contingency table, which cross-classifies fragments of the analysed corpus, and the corresponding textual units. Using Correspondence Analysis, information related to the context or to the syntactical role of textual units is generally neglected. Moreover, it is well-known that one of the critical points in performing textual data analysis is the identification of the proper textual units. Different choices have been proposed in literature (e.g. graphical forms, lemmas, textual forms, repeated segments). In any case, when texts are pre-processed for building **T**, pieces of information are destroyed. The paper aims at (partially) recovering all the lost information. Our proposal consists in a principal axes method, which enables the introduction of external knowledge both on fragments (**T** columns) and on units (**T** rows), by two additional matrices, respectively, **X** and **Z**. In particular, **Z** allows to introduce knowledge about pre-processing, or e.g. syntax, homographs, or synonymies, while **X** defines a proper partition of the fragments, as in usual correspondence analysis of aggregated lexical tables.

Riassunto

L' analisi delle corrispondenze è in genere utilizzata per identificare e rappresentare la struttura di associazione presente in una tabella lessicale **T**. **T** è una particolare tabella di contingenza che incrocia i frammenti del corpus analizzato con le corrispondenti unità testuali. Applicando l' analisi delle corrispondenze, l' informazione relativa al contesto o al ruolo sintattico delle unità testuali è ignorata. Inoltre, è ben noto che uno dei punti maggiormente critici dell' analisi dei dati testuali è l' identificazione delle unità testuali appropriate. Esistono diverse proposte in letteratura: forme grafiche, forme testuali, lemmi, segmenti ripetuti, ad esempio. In ogni caso, quando il testo è pretrattato nella costruzione di **T**, si perde informazione. Il presente lavoro ha l' obiettivo di recuperare, almeno in parte, queste informazioni perdute. Si propone, infatti, un metodo ad assi principali, che consente di introdurre informazione esterna sia sui frammenti (colonne di **T**), sia sulle unità (righe di **T**), attraverso, rispettivamente, due matrici aggiuntive **X** e **Z**. In particolare, **Z** consente di introdurre elementi relativi al pretrattamento, alla sintassi, a omografie, o sinonimie, ad esempio, mentre **X** definisce una partizione dei frammenti, come in un' analisi delle corrispondenze di tabelle aggregate.

Parole-chiave: Tabella lessicale, Pretrattamento, Analisi Fattoriale, Analisi Fattoriale non simmetrica.

1. Introduzione

Le tecniche esplorative di analisi dei dati sono utilizzate per ricercare e descrivere strutture di associazione e di dipendenza tra le caratteristiche proprie di un collettivo statistico (ad esempio, rispondenti a domande aperte in un questionario) e quelle distintive del particolare vocabolario utilizzato. La tecnica cui più comunemente si ricorre per ottenere una sintesi di

queste strutture è l' Analisi delle Corrispondenze, condotta su una particolare tabella di contingenza, detta tabella lessicale aggregata. Quest' ultima incrocia le modalità di una o più variabili categoriche con le unità testuali (Lebart et al., 1998). L' esempio più tipico riguarda l' incrocio fra le risposte fornite ad una domanda aperta ed una chiusa, in un' indagine da questionario. In questo modo, l' analisi tiene conto di informazioni relative ai frammenti di testo (nell' esempio, gli intervistati), ma non sulle unità testuali (che, nel seguito, chiameremo, per semplicità, *parole*). E', comunque, opportuno ricordare che uno dei punti più critici dell' analisi dei dati testuali riguarda proprio le decisioni relative al *vocabolario* che sarà poi l' oggetto dell' analisi. Le cosiddette "scelte di lemmatizzazione" (Bolasco, 1993) che portano all' individuazione delle parole da analizzare e, quindi, alla costruzione del vocabolario di riferimento, hanno in sé implicita una perdita di informazione, dovuta sia alla decisione di modificare (o no) la forma osservata e a come farlo, sia al forte legame del significante con il suo significato, legame che assume particolari connotazioni all' interno dello specifico campo di interesse e delle particolari esigenze di ricerca.

Nella pratica più frequente, si ricorre ad una serie di operazioni di pre-trattamento del testo, prima di arrivare alla costruzione della matrice di base dell' analisi. Tali operazioni, solitamente, consistono nel rendere i termini omogenei nel genere, nel numero e nella forma verbale (lemmatizzazione grammaticale). Nel ricorso alle cosiddette forme testuali (Bolasco, *ibidem*) ulteriori interventi, assimilabili ad operazioni di codifica, sono effettuati. Si tratta, ad esempio, di distinguere le stesse parole, se usate all' interno di strutture lessico-grammaticali o isolatamente, oppure, di fondere forme giudicate, nello specifico contesto, equivalenti, e così via. Queste operazioni preliminari costringono l' analista a trascurare, se non a distruggere, una mole d' informazione insita nel testo originario e gioco forza tralasciata in questa fase, conducendo ad un impoverimento dal punto di vista della variabilità dell' espressione. Allo stesso tempo, però, esse producono degli effetti desiderabili in termini di omogeneizzazione delle unità osservate. La stessa scelta di operare sulle singole forme grafiche, definite come sequenze di caratteri all' interno di due delimitatori (Lebart, 1981), ha in sé rischi di lemmatizzazione implicita (ancora Bolasco), da non sottovalutare.

La soggettività delle scelte che caratterizzano questa fase rende opportuno un tentativo di recupero almeno parziale di questa informazione. L'ob iettivo di questo lavoro è proporre una particolare tecnica fattoriale, basata sul ricorso a strutture informative esterne, relative sia ai frammenti, sia alle parole. I risultati dell' approccio proposto consentono di evidenziare, attraverso rappresentazioni grafiche in sottospazi fattoriali, le associazioni tra le modalità criterio introdotte sulle righe e sulle colonne della tabella lessicale.

In un precedente lavoro (Balbi, Giordano, 1999), il tema era stato affrontato all' interno di un contesto più propriamente quantitativo: il generico elemento della tabella lessicale era visto come l' intensità di risposta in corrispondenza di specifici fattori relativi alla natura delle parole e, congiuntamente, dei frammenti del *corpus*. La strategia proposta era così incentrata su di un ricorso intensivo alla regressione multipla, per comprendere i legami esistenti fra le due strutture informative esterne. In questo lavoro, si affronterà il problema nei termini di un' analisi di distribuzioni di frequenza congiunte, alla luce della partizione indotta sui frammenti di testo analizzati. L' idea è quella di realizzare un' analisi non simmetrica, nella quale l' associazione fra parole e informazioni ad esse relative viene analizzata alla luce delle diverse caratteristiche considerate nella classificazione dei frammenti. Questo comporta un approfondimento circa la metrica da adottare nelle rappresentazioni fattoriali, ma consente la visualizzazione del ruolo giocato dalle differenti parole, alla luce, ad esempio, del trattamento subito in fase di definizione preliminare del vocabolario.

Il metodo è illustrato all' interno di uno specifico problema, quello della definizione di "qualità". Oggi, infatti, questo termine è diffusamente utilizzato sia nel linguaggio comune, sia nei discorsi economici, o politici, con accenti ed implicazioni di diverso tipo. Nel presente lavoro si è cercato di analizzare l' introduzione del tema della "qualità" in tre differenti tipi di pubblicazioni, di cui è l' argomento principale, di regola già presente nel titolo. Si tratta di volumi tutti, comunque, appartenenti ad un ambito di tipo aziendale: monografie a carattere teorico, atti di convegni, manuali relativi a specifici settori di applicazione. Le informazioni relative alle unità testuale riguardano il tipo di pre-trattamento subito da ciascuna parola nei diversi testi. Il ricorso a strutture di informazione esterne appare di rilievo per perseguire l' obiettivo prefissato, ossia comprendere le differenze di vocabolario che scaturiscono dal destinatario cui lo scritto è rivolto.

2. La struttura dei dati

Il tradizionale punto di partenza per l'analisi dei dati testuali è la costruzione della tabella lessicale $\mathbf{T}_{(p\times n)}$ che incrocia le p unità testuali con gli n frammenti di testo. Con l'obiettivo di utilizzare parte dell'informazione relativa al pretrattamento operato, viene costruita una prima tabella (che indichiamo con $\mathbf{Z}_{p\times q}$), in cui descriviamo per ciascuna parola il tipo di trattamento effettuato. Il generico elemento di tale tabella contiene la frequenza con la quale ciascuna parola ha subito il q-esimo trattamento.

Consideriamo, inoltre, la tabella $\mathbf{X}_{(k\times n)}$ il cui elemento generico indica, in codifica disgiuntiva completa, l'appartenenza di ciascun frammento di testo ad un elemento della partizione indotta da una variabile categorica a k modalità (figura 1). Si ritiene, cioè, che esista un effetto derivato dall'aggregazione dei frammenti di testo secondo le modalità definite nelle righe della **X**..

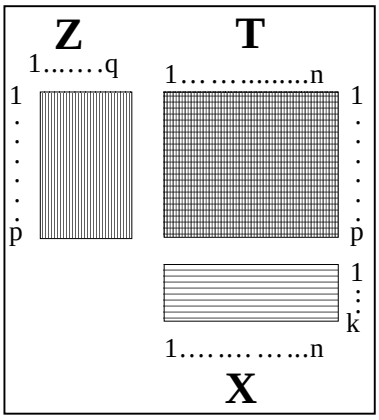

Figura 1: La struttura dei dati

Ai fini della nostra analisi, effettuiamo il prodotto matriciale **P** = **T`Z**, la matrice che ne risulta contiene in riga gli n frammenti di testo ed in colonna i q pretrattamenti utilizzati; l'elemento $p_{n,q}$ considera la frequenza con la quale un particolare trattamento è stato effettuato nel frammento di testo Dalla matrice **P** consideriamo i marginali di riga e di colonna e definiamo, rispettivamente, con ${}_c\mathbf{D}_p$ e con ${}_r\mathbf{D}_p$ la matrici quadrate che contengono sulla diagonale principale tali marginali. Dal prodotto $\mathbf{P}\ {}_c\mathbf{D}_p^{-1}$ otteniamo la matrice dei profili colonna $\mathbf{F}_{n,q}$, che ci informa del peso relativo con cui ciascun frammento di testo ha subito un operazione di pretrattamento.

A questo punto si ha interesse a rilevare in che modo l'informazione, nota a priori, sulla tipologia di frammento, può avere avuto un ruolo nella determinazione dei pesi relativi in **F**.

Effettuiamo a tale scopo un operazione di centratura della matrice **F** rispetto alle colonne e consideriamo la relazione:

$$\overline{\mathbf{F}} = \mathbf{X`B} + \mathbf{E} \qquad (1)$$

La (1) descrive il sistema di *q* regressioni dei profili **F** sul sottospazio generato dai *k* criteri (predittori qualitativi) in **X**.

Le matrici $\mathbf{B}_{(k,q)}$ ed **E** rappresentano rispettivamente la matrice dei coefficienti ed i residui dei *q* modelli. La descrizione dei profili **F** può dunque effettuarsi tramite le *k* modalità attinenti al tipo di frammento, considerando l'immagine dei profili **F** sul sottospazio generato dalle colonne della **X**, attraverso l'operatore di proiezione ortogonale $\mathbf{X`(XX`)^{-}X}$.

$$\hat{\mathbf{F}} = \mathbf{X`(XX`)X}\overline{\mathbf{F}} \qquad (2)$$

A questo punto, consideriamo il diverso numero di trattamenti che ha subito ogni frammento di testo, ed effettuiamo una ponderazione delle quantità in $\hat{\mathbf{F}}$ attraverso la matrice $_r\mathbf{D}_p$, precedentemente definita.

$$\mathbf{Q} = \mathbf{D}_p\hat{\mathbf{F}} \qquad (3)$$

La ponderazione effettuata consente di riproporzionare il contributo originario di ciascun frammento di testo secondo il criterio definito dal tipo di trattamento, tenendo conto direttamente dell'informazione relativa alla tipologia di frammento.

Il passo finale della strategia proposta consiste nella determinazione di una sintesi fattoriale degli *score*, così determinati, attraverso la risoluzione dell'equazione caratteristica:

$$\mathbf{Q`Q} = \hat{\mathbf{F}}`_r\,\mathbf{D}_p^2\hat{\mathbf{F}}_r\,u_\alpha = \mathbf{\Lambda}_\alpha u_\alpha \qquad (4)$$

La determinazione degli elementi caratteristici della decomposizione effettuata consente, tra l'altro, un'agevole interpretazione in termini geometrici e grafici delle relazioni tra le diverse entità oggetto di studio (frammenti, tipo di pretrattamento).
Può essere interessante, inoltre, considerare il ruolo svolto da ciascuna "parola" nell'analisi effettuata. A tal fine, dalla tabella **Z**, consideriamo le frequenze marginali di riga e perveniamo alla matrice dei profili di riga **R** = *r***D***z*. I profili, opportunamente centrati, possono essere rappresentati in supplementare sul piano fattoriale generato dalla (4).

3. Le parole della qualità

La tecnica proposta assume una particolare rilevanza in alcuni contesti caratterizzati da una forte eterogeneità nell' utilizzo e nel significato di alcuni vocaboli tipici. Il caso che viene qui analizzato è quello della *qualità*. Di tale parola si fa uso (ed abuso) in molteplici contesti: dai messaggi pubblicitari, ai discorsi politici, alle relazioni economico-aziendali, per citarne alcuni. A seconda dei diversi contesti, il termine *qualità* è associato o sovrapposto ad altri termini, in condizioni inammissibili in ambiti differenti (si veda Balbi, 1998, per una sua sinonimia con *affidabilità*, all' interno di uno specifico linguaggio pubblicitario). Nel presente lavoro si è deciso di circoscrivere l' analisi ad un ambiente aziendalistico, escludendo, anche, il cosiddetto controllo statistico della qualità, per orientarsi aprioristicamente ad un' accezione di *qualità totale*. Si sono, pertanto, considerati 19 volumi in lingua italiana, che si ritiene rappresentino adeguatamente il panorama oggetto di analisi. All' interno di questi testi si sono selezionati, di regola nel capitolo/paragrafo introduttivo, i periodi che definivano il punto di vista dell' Autore sul tema e, per tutti, si sono considerate duecento parole, al fine di ottenere una lunghezza identica per i diversi frammenti. Si è, quindi, proceduto ad una prima lettura

completa dei termini utilizzati, per poi operare delle selezioni, eliminando le parole strumentali (articoli, preposizioni, congiunzioni) e le parole al di sotto di una soglia di frequenza 2. Si è, quindi, proceduto ad una costruzione del vocabolario che ha visto coesistere le forme grafiche, le forme testuali e la lemmatizzazione. La tabella **T** risulta così composta dai 19 testi e dalle parole che costituiscono il vocabolario così manipolato. La matrice **Z** è, a sua volta, una tabella di contingenza che ha per elemento generico la frequenza relativa con cui un termine del vocabolario è stato sottoposto ad un determinato pre-trattamento. Circa i volumi si è considerato se questi fossero diretti ad esperti (monografie teoriche), fossero atti di convegni specialistici, o fossero rivolti ad operatori (manuali applicativi ad un determinato settore, come, ad es., piccole e medie imprese, pubblica amministrazione, scuola). **X** è, quindi, una matrice indicatrice i cui elementi sono pari a 1 se il volume appartiene alla categoria corrispondente, e 0 altrimenti.

3.1 I risultati dell'analisi fattoriale

La tabella (1) riporta la matrice **Q** degli *scores* ottenuti per i 19 frammenti di testo. I risultati della decomposizione fattoriale, condotta su tale matrice, evidenziano come l'intero contenuto informativo sia rappresentabile sulle prime due dimensioni fattoriali (si veda Tabella 2).

	forma	**lemma**	**contesto**	**radice**
fram1	-2.93	6.07	-7.53	-18.7
fram2	-10.2	-13.0	5.47	19.05
fram3	7.99	0.24	4.95	9.68
fram4	4.01	0.12	2.48	4.85
fram5	-7.71	-9.80	4.10	14.27
fram6	-7.41	-9.42	3.94	13.71
fram7	3.73	0.11	2.31	4.52
fram8	-1.80	3.73	-4.63	-11.5
fram9	4.67	0.14	2.89	5.66
fram10	-1.70	3.53	-4.39	-10.9
fram11	-1.77	3.67	-4.56	-11.3
fram12	-2.47	5.12	-6.36	-15.8
fram13	-1.42	2.95	-3.67	-9.13
fram14	-2.95	6.10	-7.58	-18.8
fram15	9.59	0.29	5.942	11.61
fram16	-2.19	4.54	-5.64	-14.0
fram17	6.99	0.21	4.33	8.47
fram18	7.76	0.23	4.80	9.39
fram19	4.26	0.13	2.63	5.15

Tabella 1: La matrice di base dell'analisi fattoriale

λ_α	%
60.8	69.0
27.3	31.0
n.a.	0.00
n.a.	0.00

Tabella 2: Autovalori e percentuale d'inerzia spiegata

L'importanza di ciascuna forma di pretrattamento utilizzata, sui primi due assi fattoriali è riportata in Tabella 3. Si evidenzia come la fonte di eterogeneità più importante è costituita dall'operazione che, qui per semplicità, abbiamo definito "*radice*" e che è consistita nel riportare alla radice comune del fonema le diverse forme grafiche con le quali esso appariva nel frammento di testo. Il trattamento definito "*forma*" ha un peso sensibilmente minore, ciò è dovuto alla circostanza che con esso si sono caratterizzati quei fonemi di cui è rimasta inalterata la forma grafica, e che essi hanno rappresentato la maggior parte dei casi. Le operazioni di lemmatizzazione ("*lemma*") e di contestualizzazione ("*contesto*") si contrappongono sul primo asse fattoriale, caratterizzandosi per la diversa tipologia di unità testuale a cui sono applicate in prevalenza. E' interessante notare come il secondo asse si caratterizzi alla stregua di un fattore di *size*, attribuendo una coordinata maggiore alla forma di trattamento più ricorrente.

	Asse 1	***Asse 2***
forma	0.07	0.88
lemma	-.32	0.45
contesto	0.34	0.16
radice	0.88	0.03

Tabella 3: Vettori unitari degli assi principali

La rappresentazione congiunta *trattamenti-frammenti di testo*, riportata in Figura 1, evidenzia come le tre tipologie di frammenti di testo si separano sul piano e come ciascun tipo si caratterizzi per i pretrattamenti effettuati, coerentemente con l'interpretazione fornita per esse rispetto a ciascun asse. In particolare, sul versante negativo del primo asse fattoriale si collocano le "*monografie*", su quello positivo i "*manuali*", mentre i frammenti "2", "5" e "6" rappresentano gli "*atti di convegno*".

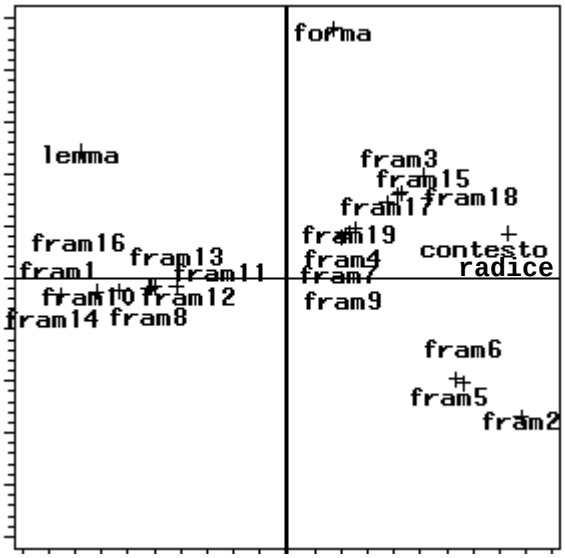

Figura 1:Rappresentazione congiunta Frammenti - Trattamenti

La rappresentazione delle unità testuali sul piano fattoriale consente di arricchire l'interpretazione, consentendo di individuare alcuni aspetti caratterizzanti della metodologia proposta. Si evidenzia come le forme grafiche che sono rimaste invariate (PIU'; MA; NON; COME; QUALITA' TOTALE) collocandosi in corrispondenza del pretrattamento "*forma*"

hanno, per lo più, caratterizzato i "*manuali*". Le "*monografie teoriche*", caratterizzate dalle operazioni di lemmatizzazione sono rappresentate in maggioranza da forme verbali (ESSERE; POTERE; SIGNIFICARE; INTENDERE) e da alcuni sostantivi tecnici (PROCESSO; PRODOTTO; TECNICA; INTERNO). Infine, gli "*atti di convegno*" manifestano con ricorrenza dei riferimenti normativi, storici, metodologici e si caratterizzano per operazioni legate al trattamento di tipo "contesto" e di tipo "radice".

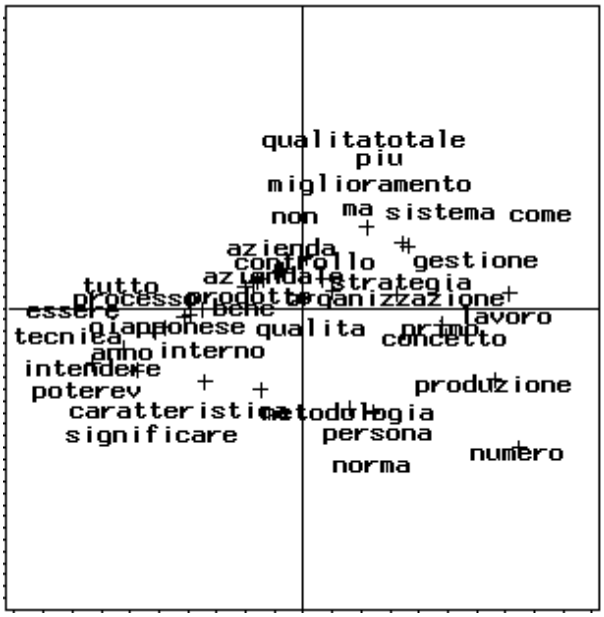

Figura 2: Rappresentazione in supplementare delle unità testuali

Riferimenti bibliografici

Balbi S., "Lo studio dei messaggi pubblicitari con l' analisi dei dati testuali", Quaderni del Dipartimento di Scienze Economiche e Statistiche, **1**, *Linguistica e Statistica: strategie di lettura* (a cura di A. Aruta Stampacchia), Dipartimento di Scienze Economiche e Statistiche, Univ. "Federico II" di Napoli, 155-171, 1998

Bolasco S. (1993). Choix de lemmatisation en vue de reconstructions syntagmatiques du texte par l' analyse des correspondance. Montpellier JADT 93. Paris: Telecom.

Gabriel K. R. (1981), Biplot display of multivariate matrices for inspection of data and diagnosis, in V. Barnett (ed.), Interpreting Multivariate Data, 147-174, Chichester, Wiley.

Giordano G., Scepi G., La progettazione della qualità attraverso l' analisi di strutture informative differenti, in Atti della XXXIX Riunione Scientifica SIS, Sorrento (in press).

Greenacre, M. (1984), Theory and Applications of Correspondence Analysis, London, Academic Press.

Lebart, L., Salem, A., Berry, L. (1998), Exploring Textual Data, Kluwer Academic Publishers, Dordrecht, The Netherlands.