doi.org/10.26398/IJAS.0029-014

A JOINT ANALYSIS OF HETEROGENEOUS INFORMATION ON ITALIAN LISTED FIRMS

Simona Balbi, Maria Spano

DiSES, University of Naples Federico II, Naples, Italy

Michelangelo Misuraca[1]

DiScAG, University of Calabria, Arcavacata di Rende, Italy

Abstract. *In recent years, the amount and variety of data available in digital format have exponentially increased. New technologies enables collecting, storing, transferring, and analysing huge amounts of data. Large digital archives, containing heterogeneous data (texts, figures, images, sounds) are easily available. The challenge is to extract useful information, by means of new methodological and computational tools, or with well-known tools used in innovative ways. Here we evaluate firms' performances, jointly analysing financial measures and management commentaries. The data structure consists of two matrices, sharing the same rows (firms), a document-term matrix and a numerical matrix. In the framework of geometric data analysis, we use a graphical approach aiming at visualising both textual descriptions and financial indices.*

Keywords*: Textual data Analysis, Geometric data analysis, Business performances*

1. INTRODUCTION

In recent years, the amount and variety of data available in digital format have exponentially increased. Let us think of the World Wide Web with its social networks, the traces left on e-commerce sites, the search engines, and so on.

New technologies enable to collect, store, transfer, and combine huge amounts of data. Therefore, an ever-increasing number of public and private institutions builds up large digital archives, containing documents, numbers, tables, images and sounds. The actual problem for data analysts is extracting useful information. One of the most stimulating challenges consists in proposing methodological tools for analysing heterogeneous data.

[1] Corresponding author: Michelangelo Misuraca, email: michelangelo.misuraca@unical.it

Here we focus our attention on a common situation, when we have both numerical data and textual descriptions. Our aim is to evaluate different strategies proposed for the joint analysis of different kinds of data.

The methodological frame has been originated by the principal component analysis with instrumental variables (Rao, 1964), which is the first reference of analyses introducing external information. This approach is also known as redundancy analysis (Van den Wollenberg, 1977).

The data structure consists of two matrices, both having the units involved into the analysis as row dimension: a matrix with textual variables (i.e., terms) and a matrix with quantitative variables. By taking into account the wide statistical literature on the analysis of two or more sets of variables, we prefer the geometric data analysis approach. The result consists in a graphical representation of the vocabulary peculiarities, with respect to the different quantitative characteristics.

The case study to be presented is based on a sample of 49 firms listed on the Italian Stock Exchange. In some countries – including Italy – the companies that want to be listed on the Stock Exchange have yearly to write a narrative business report called management commentaries (MC). The research hypothesis is that the language used by firms in the MC depends on the performances obtained by these businesses themselves.

For each selected firm we have considered the official MC presented in 2010, and some indicators commonly used in the evaluation of business performances. The influence of the chosen indicators on *corporate disclosure* was highlighted in the accounting studies' domain (Berger, 2011). Moreover, the readability of the commentaries was also considered by using a measure developed for the evaluation of documents written in Italian. We study the *corpus* of commentaries from a statistical point of view, considering different choices in metrics and weighing systems, and underlying how those choices are strictly related to different methods.

After presenting the main results in a comparative perspective, we discuss new questions and future developments.

2. THEORETICAL FRAMEWORK

Different kinds of information are often available on a given phenomenon. For example, we have verbal descriptions together with tables containing measurable characteristics. For statisticians, common practice consists in analysing numerical data, considering textual information as interesting elements for interpreting results. Wishing a deeper use of documentary information, natural language processing tools were adopted in order to transform "unstructured" data (e.g., texts) into "structured" data (e.g., numbers). In statistical literature, there are many methods developed for the analysis of two (or more) sets of variables, describing the same individuals (first reference: canonical correlation analysis, Hotelling, 1936). Here there is the additional problem that it is not easy mixing texts with other information, because of the different nature of data.

Let us consider a *matrix* **T** with *n* rows (documents) and *w* columns (terms), obtained by adopting a bag-of-words coding. In this coding scheme a document is represented as the bag (multiset) of its terms, disregarding the grammar and the context of use. The generic element t_{ij} typically represents the frequency of the *j-th* term in the *i-th* document ($i = 1, ..., n$; $j = 1, ..., w$).

Supposing to have an additional information on the documents and/or the terms, we define the $n \times p$ matrix **G** and/or the $w \times q$ matrix **H**. The matrix **G** contains some characteristics of the *n* documents, e.g., the point in time each document was written, or the main topic of the document. Similarly, the matrix **H** contains some characteristics of the *w* terms, e.g., the grammatical category (Giordano and Balbi, 2001), or the polarity in the framework of sentiment analysis (Turney, 2002).

According to Takane and Hunter (2001), the use of additional information on rows and columns of a matrix **T** can be modelled by considering:

$$\mathbf{T} = \mathbf{G}\mathbf{M}_1\mathbf{H}^\mathsf{T} + \mathbf{G}\mathbf{M}_2 + \mathbf{M}_3\mathbf{H}^\mathsf{T} + \mathbf{E} \tag{1}$$

where $\mathbf{M_1}$ $(p \times q)$, $\mathbf{M_2}$ $(n \times q)$ and $\mathbf{M_3}$ $(p \times w)$ are matrices of unknown parameters, and **E** (*n* by *w*) is a matrix of residuals. The first element concerns what can be jointly explained by **G** and **H**, the second one what can be explained by **G** but not by **H**, the third one what can be explained by **H** but not by **G**. The last element in the model pertains to what can not be explained by **G** and **H**.

In a general scheme, we also consider two metric matrices **N** and **K**, referring to the row and column sides respectively. These matrices play an essential role when it is necessary to give a different importance to the elements listed on the rows and columns of **T**. In a textual data analysis framework, it means to consider how each document and/or each term contribute to the explanation of the association structure of the data.

The parameters in the model can be estimated by minimising the residuals **E**. The least of squares (LS) estimates of $\mathbf{M_1}$, $\mathbf{M_2}$ and $\mathbf{M_3}$ are:

$$\hat{\mathbf{M}}_1 = \left(\mathbf{G}^\mathsf{T}\mathbf{N}\mathbf{G}\right)^{-1}\mathbf{G}^\mathsf{T}\mathbf{N}\mathbf{T}\mathbf{K}\mathbf{H}\left(\mathbf{G}^\mathsf{T}\mathbf{K}\mathbf{H}\right)^{-1} \tag{2a}$$

$$\hat{\mathbf{M}}_2 = \left(\mathbf{G}^\mathsf{T}\mathbf{N}\mathbf{G}\right)^{-1}\mathbf{G}^\mathsf{T}\mathbf{N}\mathbf{T}\left[\mathbf{I}\text{-}\mathbf{K}\mathbf{H}\left(\mathbf{H}^\mathsf{T}\mathbf{K}\mathbf{H}\right)^{-1}\right]\mathbf{K}\mathbf{K}^{-1} \tag{2b}$$

$$\hat{\mathbf{M}}_3=\mathbf{N}^{-1}\mathbf{N}\left[\mathbf{I}-\left(\mathbf{G}^{\mathsf{T}}\mathbf{N}\mathbf{G}\right)^{-1}\mathbf{G}^{\mathsf{T}}\mathbf{N}\right]\mathbf{T}\mathbf{K}\mathbf{H}(\mathbf{H}^{\mathsf{T}}\mathbf{K}\mathbf{H})^{-1} \tag{2c}$$

The estimates in (2a), (2b), and (2c) can be rewritten in terms of orthogonal projectors. The decomposition of **T**, according to (1) is:

$$\begin{aligned}\mathbf{T}=&\mathbf{P}_{\mathbf{G|N}}\mathbf{T}\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}+\mathbf{P}_{\mathbf{G|N}}\mathbf{T}\left(\mathbf{I}-\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}\right)\mathbf{K}\mathbf{K}^{-1}+\mathbf{N}^{-1}\mathbf{N}\left(\mathbf{I}-\mathbf{P}_{\mathbf{G|N}}\right)\mathbf{T}\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}+\\&+\left[\mathbf{T}-\mathbf{P}_{\mathbf{G|N}}\mathbf{T}\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}-\mathbf{P}_{\mathbf{G|N}}\mathbf{T}\left(\mathbf{I}-\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}\right)\mathbf{K}\mathbf{K}^{-1}-\mathbf{N}^{-1}\mathbf{N}\left(\mathbf{I}-\mathbf{P}_{\mathbf{G|N}}\right)\mathbf{T}\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}\,]\right].\end{aligned} \tag{3}$$

where $\mathbf{T}=\mathbf{P}_{\mathbf{G|N}}=\mathbf{N}^{1/2}\mathbf{G}(\mathbf{G}^{\mathbf{T}}\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^{\mathbf{T}}\mathbf{N}^{1/2})$ and $\mathbf{P}_{\mathbf{H|K}}=\mathbf{K}^{1/2}\mathbf{H}(\mathbf{H}^{\mathbf{T}}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{K}^{1/2}$. When **N** and **K** are both non-singular the decomposition is reduced to:

$$\mathbf{T}=\mathbf{P}_{\mathbf{G|N}}\mathbf{T}\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}+\mathbf{P}_{\mathbf{G|N}}\mathbf{T}\left(\mathbf{I}-\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}\right)+\left(\mathbf{I}-\mathbf{P}_{\mathbf{G|N}}\right)\mathbf{T}\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}+\left(\mathbf{I}-\mathbf{P}_{\mathbf{G|N}}\right)\mathbf{T}\left(\mathbf{I}-\mathbf{P}^{\mathsf{T}}{}_{\mathbf{H|K}}\right) \tag{4}$$

In order to analyse one or more terms in (4), in a geometric data analysis framework, we perform a singular value decomposition (SVD). Naming **A** the generic term we have:

$$\begin{gathered}\mathbf{A}=\mathbf{U}\Lambda\mathbf{V}^{\mathsf{T}}\\ \mathbf{U}^{\mathsf{T}}\mathbf{U}=\mathbf{V}^{\mathsf{T}}\mathbf{V}=\mathbf{I}.\end{gathered} \tag{5}$$

If we want to give different weights to the elements it is possible to introduce different orthonormalising constraints, with $\mathbf{U}^{\mathbf{T}}\mathbf{N}^{-1}\mathbf{U}=\mathbf{V}^{\mathbf{T}}\mathbf{K}^{-1}\mathbf{V}=\mathbf{I}$. Dealing with textual data, for example, we can be interested in taking into account some linguistic characteristics, as the grammatical category (nouns, verbs, articles, and so on).

In this context, the aim of SVD is to construct a lower dimensional space reflecting the semantic structures in the data. In information retrieval, SVD is the algebraic basis of latent semantic indexing (Deerwester et al., 1990).

If we are interested in explaining how the additional information influence both documents and terms, e.g., we will focus only on the first term in (4), whereas the last term will be analysed if we want to consider the residual effect of the additional information on both sides.

In this general frame it is possible to subsume some well-known techniques, by choosing different additional information as well as different constraints.

If **G=I**, **H=I**, **N=**(1/n)**I** and **K=I**, we consider no external information on documents and terms, and at the same time we give more importance to longer documents and to the most used terms. In this case **A=T* (T*** is centered) and we perform a principal component analysis. If **G=I**, **H=I**, $\mathbf{N}=\mathbf{D}_{\mathbf{n}}$ and $\mathbf{K}=\mathbf{D}_{\mathbf{w}}$, where $\mathbf{D}_{\mathbf{n}}$

and $\mathbf{D_w}$ are diagonal matrices with the marginal distributions of the rows and columns of $\mathbf{A}=\mathbf{T}/f_{..}$, with $f_{..}= \sum_{i=1,\dots,n} \sum_{j=1,\dots,w} f_{iw}$, we still avoid considering external information on documents and terms. In graphical displays, we give the same importance to shorter and longer documents as well as to rare and most used terms. We perform a correspondence analysis on the matrix $\mathbf{A}=\mathbf{T}/f_{..}$.

An alternative solution for introducing information on documents is given by correspondence analysis on aggregated tables, i.e., grouping the documents according to some common characteristics. Typical examples are given by age or gender in analysing open questions in surveys (Lebart et al., 1998; Becue and Pagès, 2015).

If **G** is a matrix of *p* variables observed on the *n* documents and **H=I**, we consider additional information only on the documents. The decomposition of **T** in (4) lessens to:

$$\mathbf{T} = \mathbf{P_{G|N}}\mathbf{T} + \left(\mathbf{I} - \mathbf{P_{G|N}}\right)\mathbf{T} \tag{6}$$

When **N**=(1/n)**I** and **K=I**, and **T** is standardised, the analysis of the first term in (6) is equal to the principal component analysis onto a reference subspace (PCAR, D'Ambra and Lauro, 1992). When instead $\mathbf{N}=\mathbf{D_n}$ and $\mathbf{K}=\mathbf{D_w}$, the analysis of the first term in (6) is equal to canonical correspondence analysis (CCA, Ter Braak, 1986). If **H** is a matrix of *q* variables observed on the *k* terms and **G=I**, we consider additional information only on the terms. By choosing a different metric we again refer to PCAR or to CCA on the table $\mathbf{T^T}$.

3. DATA STRUCTURE

A *management commentary* (MC) is a narrative yearly business report. It is a mandatory document in some countries – like in Italy – for all the companies that want to be listed on the Stock Exchange.

According to the recommendations of the International Accounting Standards Board (IASB), an MC is an essential annex to the financial statements that aims at presenting the management's view on the budgetary situation, the financial performances and the cash flows of a company. The MC helps the stakeholders in evaluating the outlook of a company and its general strengths and weaknesses, as well as the success of management strategies in achieving the proposed goals. Each MC usually presents the following basic information:

- the nature of business;
- the management's goals and the strategies for achieving these goals;
- significant resources, risks and relationships;

– results of operations and future scenarios;
– measures and indicators used for evaluating business performances.

In this paper, we have considered the management commentaries of 49 firms listed on the Italian Stock Exchange (*Borsa Italiana Spa*). The reference year is 2010.

In order to represent the different economic sectors in which companies are classified, we have extracted the sample by using a quota sampling design. We decide to exclude financial companies because their MC are subject to specific law regulations, which could be very different from those of non-financial companies.

The commentaries were downloaded from the official website of *Borsa Italiana* (http://www.borsaitaliana.it).

We focus particularly on the section usually named outlook, which is common to all management commentaries. The *corpus* consists of 20529 tokens and of 4262 types. Pre-treatment procedures were performed by using TalTac (Bolasco, 2012), one of the most widely used software for the textual analysis of documents written in Italian. Having normalised the trivial cases of lexical ambiguity, it was possible to clean up texts from empty terms (e.g., conjunction, articles, and adverbs) and from rare terms, with a number of occurrences less than five.

For each company several numerical variables can be considered. In the following, the MC readability as well as some performance indicators were used.

In order to evaluate the readability of the MC's outlook section, we use the GULPEASE index (Lucisano and Piemontese, 1988). This index is a careful and thoughtful review of earlier readability indices, like the Flesch index proposed in the 1940s by Rudolf Flesch for American English, and the Gunning Fog index (Gunning, 1952), adapted to Italian language. Readability of an Italian text is measured by applying the following formula:

$$Readability = 89 - 10^{-1} Lp + 3Fr \qquad (7)$$

where *Lp* is the ratio of the number of letters and the number of terms (in percentage) and *Fr* is the ratio of the number of sentences and the number of terms (in percentage). *Readability* values vary in a range of 0-100. For readers with an elementary education, texts are easy to read when the index is above 80; for those with a middle-level education, texts are easy to read when the index is above 60. For readers with a high-level education, texts are easy to read when the index is above 40.

The performance indicators taken into account are the *audit firm size*, the *profitability*, the *leverage* and the *firm size*. Several studies in the accounting domain proved that these indicators influence corporate disclosure in different ways.

The *audit firm size* considers the importance of the auditor company chosen by each firm (Firth, 1979; Healy and Palepu, 2001). As in previous studies, we dichotomise the auditors companies with respect to their importance in *BIG 4* (i.e., PriceWaterhouseCoopers, Ernest&Young, Deloitte&Touche and KPMG) and *OTHER AUDITORS.*

The *profitability* evaluates the ability of a firm in producing profits (Courtis, 1986; Wallace et al., 1994). We calculate profitability as the ratio between pre-tax earnings and total sales in the reporting year. The *leverage* is one of the most important indicators in Corporate Finance (Ahmed and Courtis, 1999). We consider the leverage of a firm as the ratio of the total long-term debt and the equity at the end of the reporting year (Wallace et al., 1994).

The influence of firm size on corporate disclosure has been stated in several studies (e.g., Lang and Lundholm, 1993). In our analysis we consider *market capitalisation* as a proxy of the firm size (Gabaix and Landier, 2008). Market capitalisation was obtained by multiplying the outstanding shares of the firm by the current market price of one share.

4. COMPARING PCAR AND CCA: A CASE STUDY

In the case study we present, the research hypothesis – according to a well established literature (Ahmed and Courtis, 1999) – is that the language used by firms in the MC depends on the performances obtained by the firms themselves.

We consider a 49 (firms) by 371 (terms) matrix. The other characteristics of the firms (the readability measure and the performance indicators) are organised in a matrix **G**, with 49 rows and 6 columns. Since we do not have additional information on the vocabulary, we set **H=I**. As seen before, if **N=I** and **K=I**, the analysis of **T** in the space spanned by **G** is equal to a PCAR. Differently, if $\mathbf{N}=\mathbf{D}_n$ and $\mathbf{K}=\mathbf{D}_w$, the analysis of **T** in the space spanned by **G** is equal to a CCA.

The motivations in choosing one of the two methods are related to the aim of the analysis. In both cases we are exploring the dependence of the textual information by some quantitative variables. The main differences of the two approaches concern a different view of the data types, a different metric for documents and terms, and a different centring procedure.

In a PCA viewpoint, documents are the cases and terms are the variables: the general element of the analysed matrix is the intensity of the use of a term in a document. The use of usual Euclidean metric gives more importance to longer documents and to the most used terms. Concerning the centring, we consider the deviation with respect to the average for each term. From a CA viewpoint, the terms are considered the categories of the linguistic variable *vocabulary* and the different documents are the categories of the variable *corpus*. The use of the chi-square metric normalises the importance of documents and terms, so that the effect of document length and term frequency is dampened. The centring takes into account the distributional independence hypothesis of *vocabulary* and *corpus* in a chi-square perspective.

In the framework of constraint analyses, in PCAR and CCA we introduce additional information on the different documents. PCAR is suitable if our aim is to analyse the strength in using the different terms and their correlations with respect to the quantitative information, while CCA is useful if we are interested in analysing the variability of the language with respect to the quantitative information. In the following, the results of the different analyses are presented.

4.1 PCAR RESULTS

The first factorial plane obtained by performing the PCAR explains about the 70.0% of the constrained total inertia (axis 1: 39.05%, axis 2: 30.89%). On this plane, it is possible to represent either the firms or the terms. This means to highlight the similarities among the different firms, as well as the use of the different terms in the commentaries, both in the space spanned by **G**. At the same time, it is possible to project the columns of **G** as supplementary variables, in order to improve interpretation.

Figure 1 provides a representation of *readability*, *profitability*, *leverage*, *capitalisation* and *auditor size*.

In Figure 2 and Figure 3 the 49 firms and the MC vocabulary are represented, respectively.

The firms that are supported by the *BIG 4* in the drafting of MC seem to focus their attention particularly to their core business. The firms supported by the other auditors emphasise the obtained results and the future developments of the business. The firms with higher *readability*, higher *profitability* and higher *capitalisation*, speak concretely of their financial results without fear of thorny issues, such as the crisis or the duties to sustain. The firms with a higher *leverage* focus their attention on the future programs, and the possible scenarios of economic development.

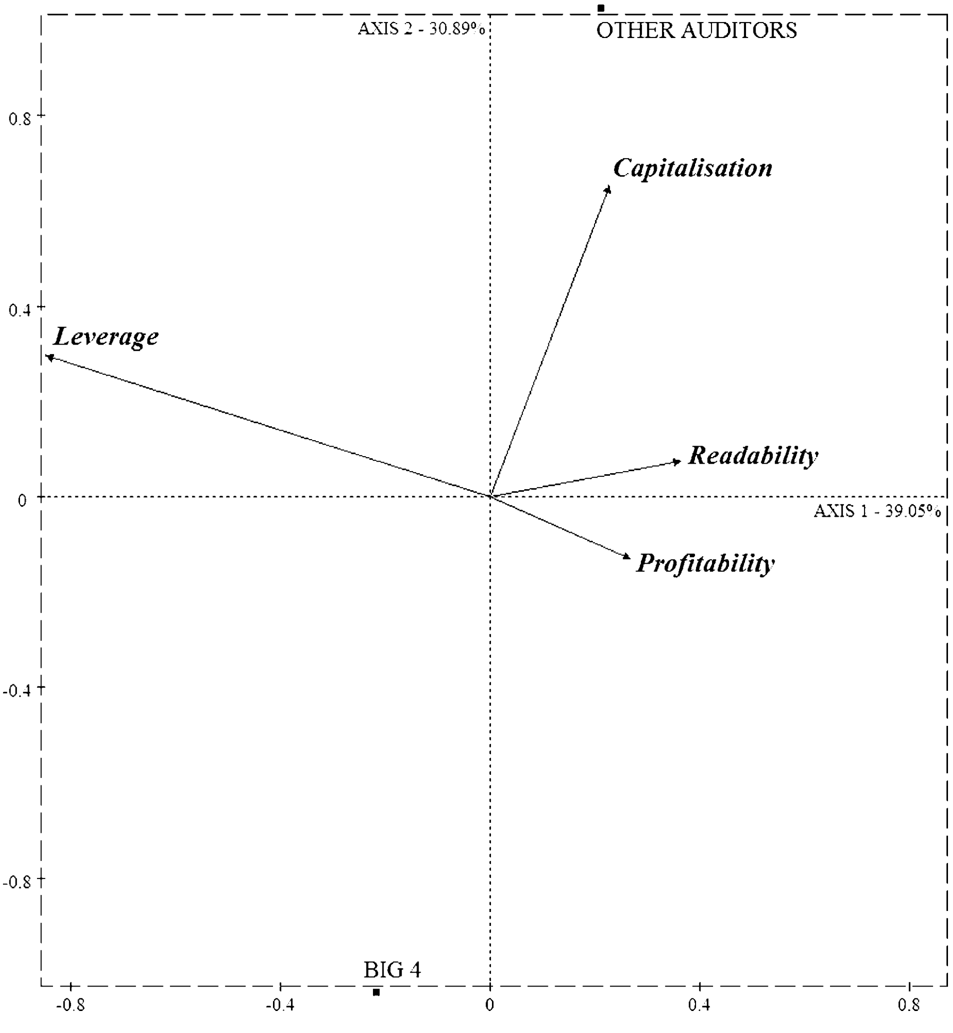

Figure 1: Indicators in PCAR first factorial plane

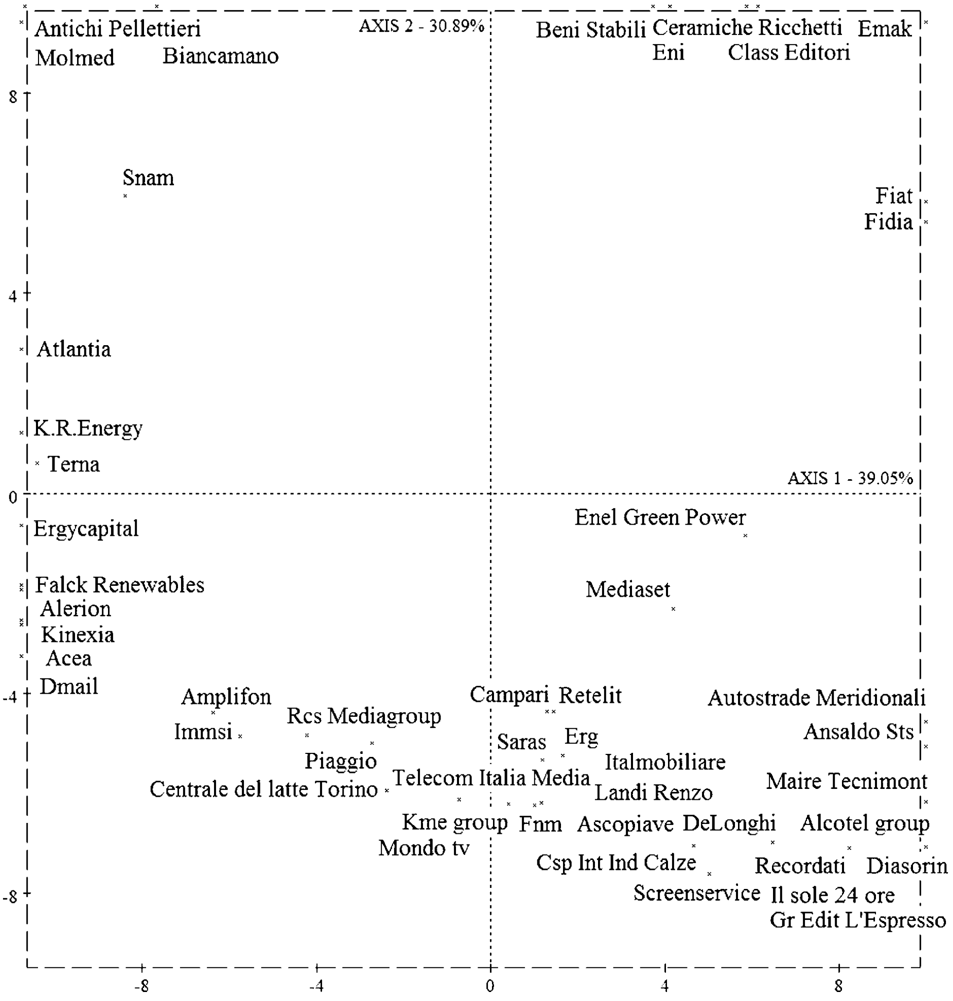

Figure 2: Firms in PCAR first factorial plane

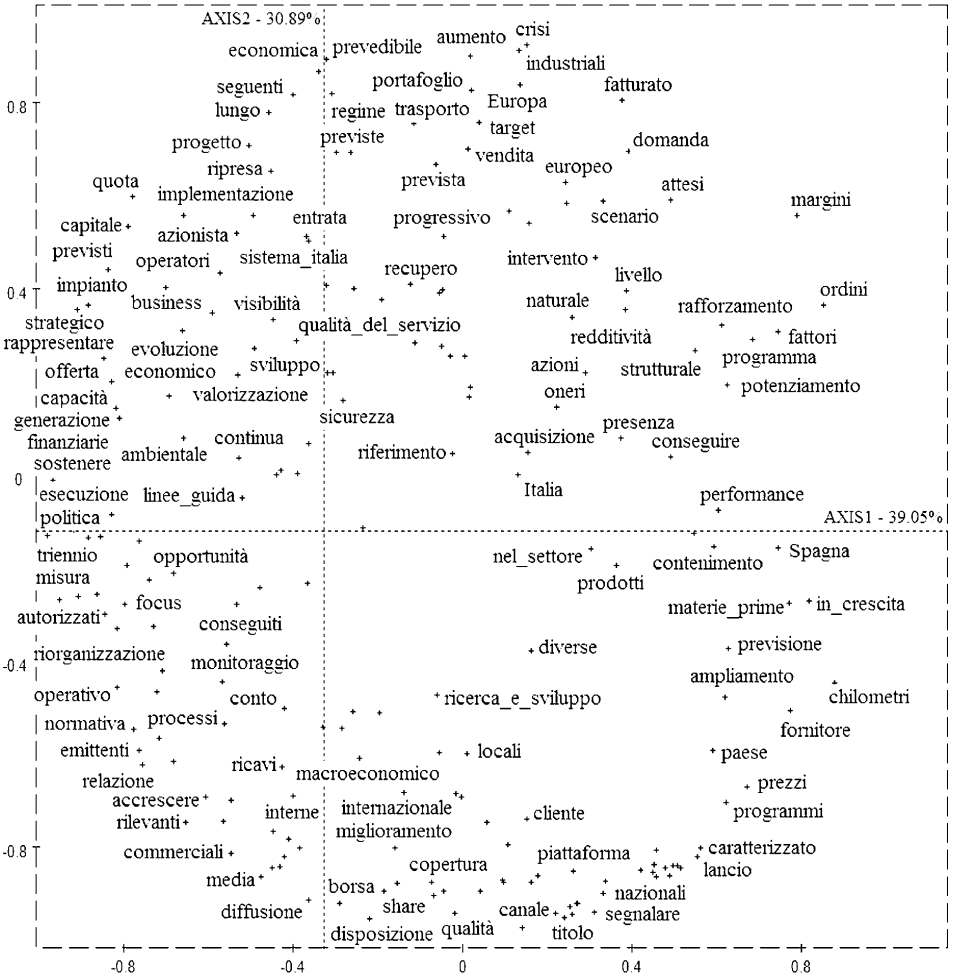

Figure 3: Terms in PCAR first factorial plane

4.2 CCA RESULTS

In a different fashion, CCA highlights the lexical similarities as well as the vocabulary in the space spanned by **G**, by considering as metrics on the two sides the length of the outlook sections and the number of occurrences of each term belonging to the vocabulary. In this case the first factorial plane explains in this case the 55.5% of the constrained total variability (axis 1: 34.65%, axis 2: 20.83%).

Figure 4 represents the different firms together with the *readability* and the performance indicators. Figure 5 instead provides a representation of the terms used by the firms in the MC.

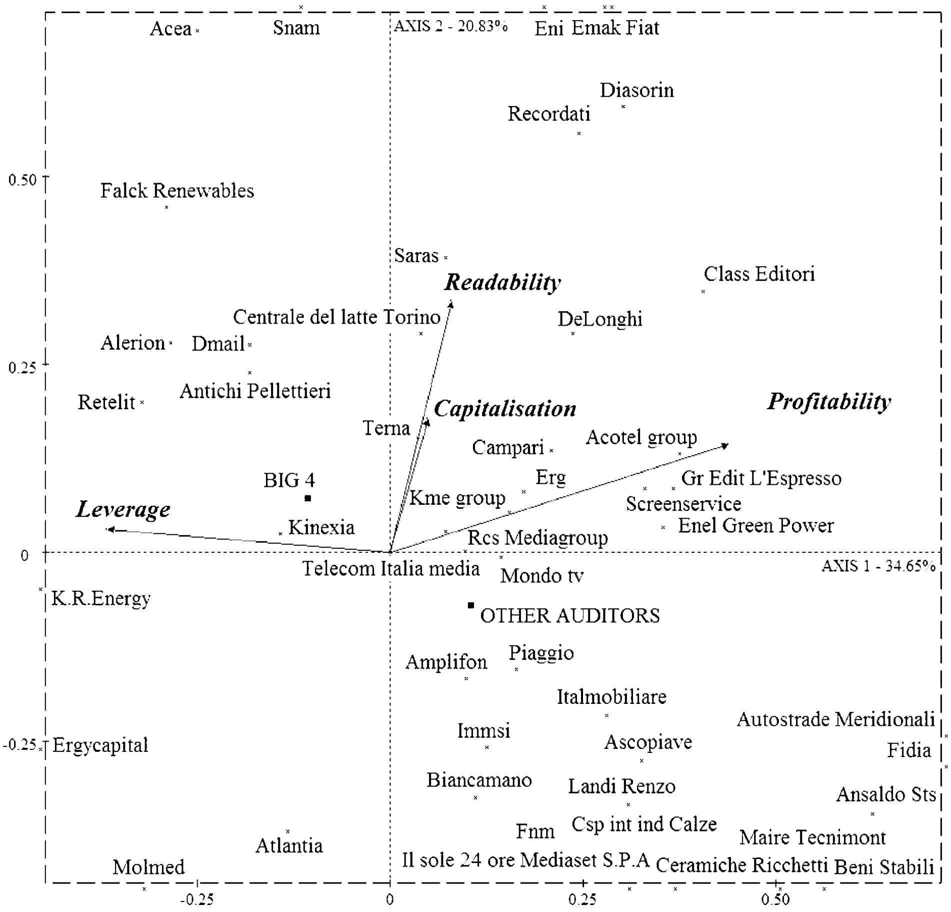

Figure 4: Firms and indicators in CCA first factorial plane

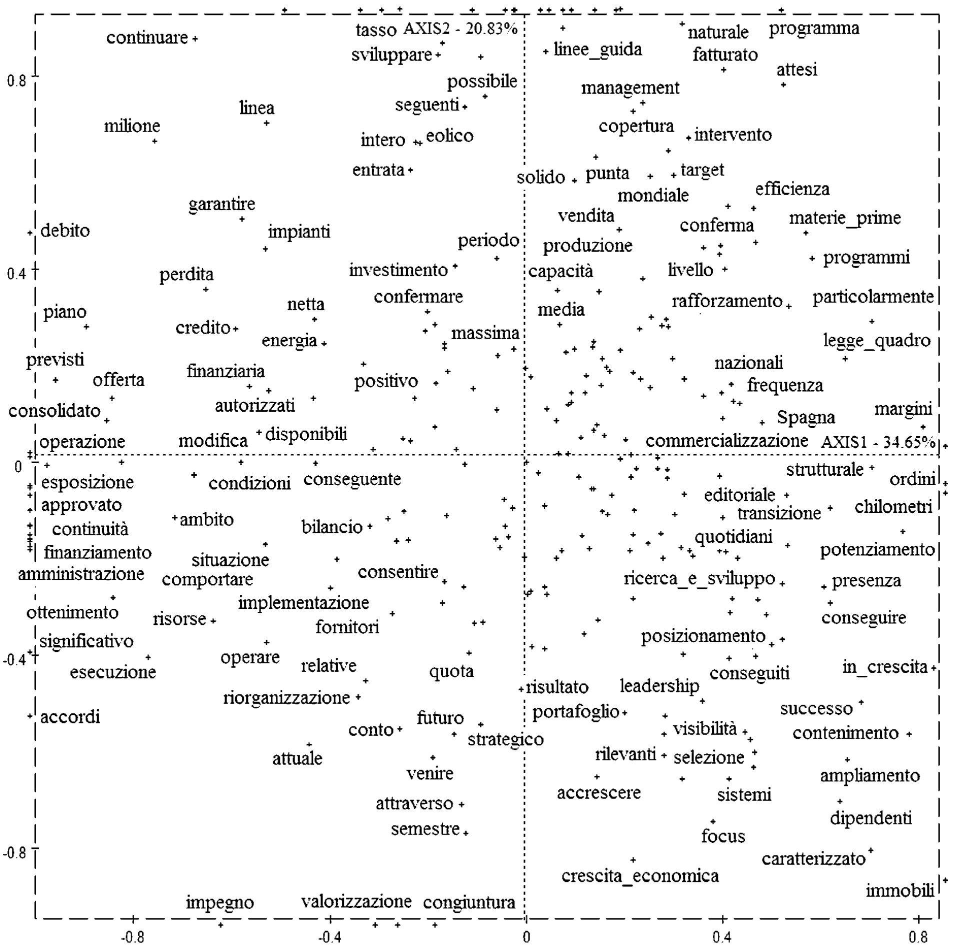

Figure 5: Terms in CCA first factorial plane

The first factorial axis is positively correlated with *profitability* and negatively correlated with the *leverage*, while the second factorial axis is positively correlated with *readability* and – to a lesser extent – with *capitalisation*.

It is interesting to notice that the relations among the variables are quite similar to the ones described in the previous analysis. The *readability* shows a higher correlation with *capitalisation* rather than *profitability*, as shown above.

The firms with higher *readability*, higher *profitability* and lower *leverage*, focus their attention on future programs and discuss their expectations of improvements. In contrast, firms with lower *readability*, higher *profitability* and lower *leverage* pay more attention to the economic aspects. Firms with lower *readability*, higher *leverage* and lower *profitability* focus their attention on debts, risks factors and internal factors. The lexical structure is coherent with these variables: debts and risks for high *leverage*, and programs and expectations for *profitability*.

4.3 A COMPARISON

There are some interesting issues to be considered in the choice between applying a CCA or a PCAR. As a matter of fact, the two methods are useful in showing some different aspects of the analysed phenomenon.

In our case study, the second factors of the two analyses have different meanings. While in PCAR the second axis can be read in terms of opposition between *BIG 4*'s clients and *OTHER AUDITORS*'s clients, in CCA, the second axis can be interpreted in terms of performance (although the opposition *BIG4* and *OTHER AUDITORS* remains). Furthermore, the power of synthesis of PCAR is higher than CCA. The inertia explained by PCAR is 70.0%, while the inertia explained by CCA is 55.5%, when we consider the first factorial plane.

On the other way round, CCA, being a correspondence analysis, enables the joint plot of terms, firms and indicators (with the usual warnings of CA). PCAR, being a principal component analysis, does not allow joint plots, due to the different metrics in the row and column spaces. Indicators are represented as supplementary points, therefore they are useful in interpreting the factorial maps.

5. DISCUSSION

The complexity of natural language as a statistical phenomenon should require all the available information about the context in which the documents were produced, as well as the subjects the documents themselves are referred to directly or indirectly. In the frame of geometric data analysis, this *meta-information* can be expressed as one or more characteristics, and it can be taken into account in several ways. The most trivial way is to project these characteristics as supplementary variables, but it is more significant when they play an active role in the analysis. As discussed above, an interesting solution is to consider a CCA approach (Ginesti et al., 2012). This allows to decompose the linguistic variability and to highlight the latent semantic structures of a collection, by taking or not taking into account the effect of the different characteristics themselves.

In the specific context of textual data analysis, the use of several characteristics at the same time – in a viewpoint of constrained analyses – has not been deeply explored in the past. This is even truer for quantitative information, that can effectively support the interpretation of textual data.

Moreover, the use of different metrics provides different perspectives in the study of the association structure of data. The use of metrics and weights for rows and columns, i.e., documents and terms, have been explored and discussed in several contributions (e.g., Balbi and Misuraca, 2005). Analogously, the choice of a metric can influence the representation of the phenomenon and its interpretation also in the frame of constrained analyses. Usually, the use of an Euclidean metric or a chi-square metric in the domain of textual data depends on the interests of the researchers. Nevertheless, sometimes it depends also on the domain in which different proposals were originally developed. If we have a text mining standpoint, such as in information retrieval, the role of longer documents and most used terms is stressed because the aim is to satisfy a specific informative need. In a different perspective, more proper to textual data analysis and in such a way to the *French* approach to data analysis, if we have an exploratory standpoint we want to give the same importance to each document in the collection. On the other hand, it is interesting to consider both common and rare terms. These latter can really discriminate between the different group of documents or between the different emerging topics. The use of constrains and of numerical data can change these clear statements.

In which way does a "numerical" projector explains or biases the relations underlying a document-term matrix? It could be interesting to go deeply into the inner nature of textual data. In literature the use of some techniques like correspondence analysis is commonly accepted under the hypothesis that the values in a document-term matrix have to be seen as joint frequencies, and not as intensities of *terms-variables*. A different viewpoint can open new questions and perspectives.

REFERENCES

Ahmed, K. and Courtis, J.K. (1999). Associations between corporate characteristics and disclosure levels in annual reports: A meta-analysis. In *British Accounting Review*. 31: 35-61.

Balbi, S. and Giordano, G. (2001) A factorial technique for analyzing textual data with external information. In S. Borra, R. Rocci, M. Vichi, and M. Schader, editors, *Advances in Classification and Data Analysis*. Springer-Verlag, Berlin-Heidelberg: 169-176.

Balbi, S. and Misuraca, M. (2005), Visualization techniques in non symmetrical relationships. In S. Sirmakessis, editors, *Knowledge Mining (Studies in Fuzziness and Soft Computing)*, Springer-Verlag, Heidelberg: 23-29.

Bécue-Bertaut, M. and Pagès, J. (2015). Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. In *Advances in Data Analysis and Classification*. 9(2): 125-142.

Berger, P.G. (2011). Challenges and opportunities in disclosure research - A discussion of the financial reporting environment: Review of the recent literature. In *Journal of Accounting and Economics*. 51(1-2): 204-218.

Bolasco, S. (2012). Introduction to the automatic analysis of textual data via a case study. S*tatistica Applicata*. 21(1): 9-21.

D'Ambra, L. and Lauro, N.C. (1992). Non symmetrical exploratory data analysis. In S*tatistica Applicata*. 4(4): 511-529.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990). Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*. 6: 391-407.

Firth, M. (1979). The impact of size, stock market listing, and auditors on voluntary disclosure in annual corporate reports. In *Accounting and Business Research*. 9(36): 273-80.

Gabaix, X. and Landier, A. (2008). Why has CEO pay increased so much?. In *Quarterly Journal of Economics.* 123: 49-100.

Ginesti, G., Maffei, M., Spano, M. and Triunfo, N. (2012). A textual processing approach for analysing the narrative disclosures in corporate reports. In *Proceedings of the 7th International Conference Accounting and Management Information System AMIS 2012*. Editura ASE, Bucharest: 1339-1352.

Gunning, R. (1952). *The Technique of Clear Writing.* McGraw-Hill, New York.

Healy, P., Hutton, A. and Palepu, K. (1999). Stock performance and intermediation changes surrounding sustained increases in disclosure. In *Contemporary Accounting Research.* 16: 485-520.

Hotelling, H. (1936). Relations between two sets of variants. *Biometrika*. 28: 321-377.

Lang, M. and Lundholm, R. (1993). Cross-sectional determinants of analysts ratings of corporate disclosures. In *Journal of Accounting Research.* 31: 246-271.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*, Volume 4, Kluwer Academic Publisher, Dordrecht.

Lucisano, P. and Piemontese, M.E. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. In *Scuola e città*. 39: 110-124.

Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya (series A)*. 26: 329-358.

Takane, Y. and Hunter, M.A. (2001). Constrained principal component analysis: A comprehensive theory. In *Applicable Algebra in Engineering, Communication, and Computing*. 12: 391-419.

Ter Braak, C.J.F. (1986). Canonical correspondence analysis: A new eigenvector technique for a multivariate direct gradient analysis. In *Ecology*. 67: 1167-1179.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics*: 417-424.

Van den Wollenberg, A.L. (1977). Redundancy analysis. An alternative for canonical correlation analysis. In *Psychometrika*. 42(2): 207-219.

Wallace, R.S.O., Naser, K. and Mora, A. (1994). The relationship between the comprehensiveness of corporate annual reports and firm characteristics in Spain. In *Accounting and Business Research*. 25(97): 41-53.