Advances in Data Analysis

Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks

Birkhäuser
Contents

Preface .................................................. XIII
List of Contributors ....................................... XV
List of Tables ............................................ XIX
List of Figures ........................................... XXI

Part I Data Mining and Text Mining

1 Assessing the Stability of Supplementary Elements on Principal Axes Maps Through Bootstrap Resampling. Contribution to Interpretation in Textual Analysis
Ramón Alvarez-Esteban, Olga Valencia, and Mónica Bécue-Bertaut

1.1 Introduction ......................................... 3
1.2 Data .................................................. 3
1.3 Methodology .......................................... 4
1.4 Results .............................................. 4
1.4.1 CA results ......................................... 5
1.4.2 Stability .......................................... 6
1.5 Conclusion ........................................... 9
References ............................................. 10

2 A Doubly Projected Analysis for Lexical Tables
Simona Bolbă and Michelangelo Miorinca

2.1 Introduction ......................................... 13
2.2 Some methodological recall ........................... 13
2.2.1 Constrained principal component analysis ............ 14
2.2.2 Principal component analysis onto a reference subspace ................ 14
2.3 Basic concepts and data structure .................... 15
2.4 A doubly projected analysis .......................... 16
2.5 The Italian academic programs: A study on skills and competences supply ........................................ 16
References ............................................. 18
Assessing the Stability of Supplementary Elements on Principal Axes Maps Through Bootstrap Resampling. Contribution to Interpretation in Textual Analysis

Ramon Alvarez-Esteban¹, Olga Valencia², and Mónica Bécue-Bertaut³

¹ Department of Economics and Statistics, University of León, Spain (e-mail: ramon.alvarez@unileon.es)
² Department of Applied Economics, University of Burgos, Burgos, Spain (e-mail: oval@ubu.es)
³ Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain (e-mail: monica.becue@upc.edu)

Abstract: Bootstrap resampling is commonly used to assess the stability of the configurations issued from principal axes methods. In the case of textual analysis, the interpretation is usually supported by the characteristics of the individuals, used as supplementary variables. To assess the stability of these variables gives information about the global structure stability.

An example issued from a wine guide illustrates the interest of computing confidence regions for supplementary categorical or quantitative variables in correspondence analysis applied to lexical tables.

Keywords and phrases: Correspondence analysis, bootstrap, textual analysis, free-text comments

1.1 Introduction

Bootstrap resampling has shown its potentiality to assess the stability of the configurations issued from principal axes methods. It allows for computing confidence regions for the elements represented on the principal subspaces (Efron and Tibshirani, 1993; Lebart et al., 2006). In many cases, the supplementary rows and/or columns provide essential information to interpret the results. In textual studies, when correspondence analysis (CA) is performed on a lexical table crossing individuals and words, the interpretation is usually supported by the characteristics of the individuals, used as supplementary variables. To assess the stability of these variables gives information about global structure stability.

Section 1.2 presents the data. Section 1.3 reviews the principles of bootstrap, and Section 1.4 offers some results obtained in the example data. Finally, Section 1.5 concludes with some remarks.
1.2 Data

Wine tasting is becoming an increasing domain for textual data analysis. The wine guide El Mundo (El Mundo, 2005) has analysed 522 wines from 'Castile and Leon'. This region (94.273 km²) is located in the northwest of Spain and comprises five AOC designations (Bierzo, Cigales, Ribera del Duero, Rueda, and Toro).

Here, we only focus on the 364 red wines. Every wine is described by free-text tasting notes and complementary information such as score (between 70 and 97), price, type of grape, vintage, etc. (ten Kleij and Musters, 2003) (Table 1.1).

<table>
<thead>
<tr>
<th>Table 1.1. Free-text tasting notes example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wine 30 Tares P3-2001 premium. Score = 91.</strong></td>
</tr>
<tr>
<td>A lot of ‘terroir’ stands out in this great red wine bouquet; hint of minerals, silex, slate, warm roasted character with a contrast of damp soil and much ripe fruit, concentrated, fatty finish on the palate, impressive viscosity on the tongue, again, flavours of damp soil and minerals in the lengthy end.</td>
</tr>
</tbody>
</table>

A lemmatization step has been carried out (Labbé, 1990; Muller, 1977-1992). Then the nouns, adjectives, verbs, and adverbs have been selected and, among these categories, only the words used at least eight times in the whole of the tasting notes are kept. Thus, the resulting lexical table crosses 364 wines (rows) and 222 words (columns).

1.3 Methodology

To assess the stability of the configurations issued from CA, partial or total bootstrap can be considered. In the former case, the principal subspace issued from the analysis performed on the original table is considered as a reference space and the rows or columns of the replicated tables are considered as supplementary elements. In the latter case, a new analysis is performed on every replicated table and the resulting configurations are compared (Lebart, 2004). In this work, we only focus on partial bootstrap.

In the following, we use the terms of the example. Thus, the statistical units (rows) refer to the wines, the active columns represent the words, and the supplementary columns correspond to the characteristics of the individuals (quantitative or categorical).

One basic principle of bootstrap consists in reproducing the process that is used to extract the random sample from the population, but using the distribution of the observed sample as an approximate distribution of the parent population (Lebart, 2006). In our case, the wine sample selection does not follow any random method, but is explicitly chosen by the expert owing to its qualities. Thus, no actual sampling error exists. Nevertheless, bootstrap resampling can be performed, by means of drawing with replacement a sample of size 364 out of the initial wine sample. It allows for studying the stability of the results facing perturbations in the wine choices by the expert.

The replicated tables have the same columns (words) as the original table (although the word frequencies can be different) and 364 bootstrapped rows. For a particular replication, some wines may not appear whereas others may be present more than once. This step is repeated B times (in our case B = 500); from every B bootstrap sample, a replicated wines x words table is built up. At every stage, the margins can differ from the original table margins. Nevertheless, as usual in CA, the latter are used as reference to compute the coordinates of rows and columns of the replicated table, considered as supplementary elements.

Depending on the replication, the coordinates of the wines remain constant, but the coordinates of the columns vary. We can compute these coordinates for the active and supplementary columns (frequency, quantitative or categorical) and the confidence regions (Lebart, 2006; Beran and Srivastava, 1985).

1.4 Results

1.4.1 CA results

Table 1.2 shows the highest five eigenvalues as well as the proportion of total inertia that they explain.

<table>
<thead>
<tr>
<th>Table 1.2. Eigenvalues and proportion of inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axes</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

As usual in a sparse table, the first eigenvalues of the CA express a very small part of the total inertia (Lebart et al., 1998, pp. 120-126).

Despite the low percentage of inertia explained by the first axis (2.046%), the high correlation between the initial score and the first axis of CA (0.70) shows that the main dimension induced by the words expresses the score, at least for a large amount. Thus, we interpret the first axis as a score level axis (Figure 1.1).

The wines with the highest scores have positive coordinates while the wines with lowest scores have negative coordinates. On Figure 1.1, the wines located on the right have a higher score than 88 whereas the wines located on the left of the vertical show a lower score than 82.

Furthermore, to make the relationship precise between the first axis (and eventually, the second axis) the values of the score are grouped into six categories (or score levels) and projected onto the first principal plane (Figure 1.2). Except for the lower score level (level 1), the categories follow the natural order along the first axis.

The information given by the relationship between the score and the first axis allows for disclosing the meaning of the words in the context of a wine guide. For example,
concerning the words related to hedonic features, the first axis contrasts words such as impressive, fatty, nutty, gun powder, and modern on the right, with amiable, easy, traditional, consistency, and young on the left (Figure 1.3). The latter words, albeit positive in current language, present here a negative reading. We are able to assert this remark thanks to the relationship between the score and the first axis.

1.4.2 Stability

As the interpretation mainly relies on the supplementary columns, we have to combine the study of the stability of the words and the supplementary variables by means of the bootstrap procedure. Here, we favor the latter. To address this problem, 500 bootstrap resamplings on the 364 wines have been performed. For each replicated table, the coordinates of each score category are computed using the CA transition formula.

Table 1.3 shows the means and the standard deviations of the score levels. A high value of the standard deviation of the coordinates of the lower category is observed (only five wines with the lowest scores)

Figure 1.4 shows the confidence regions of every score level. The highest score levels (6: score > 89, and 5: score 86–88) present confidence regions that do not overlap with the others. On the contrary, the confidence region, as well as the high standard deviation of the lower score level on the first principal plane, shows that the first category does not hold any relationship with the first two axes.

Repeating the score as a quantitative variable, Table 1.4 shows that its correlation with the first original axis varies between 0.63 and 0.78 among the replicated tables, presenting a low deviation standard. The interpretation of the first axis as a score level axis is stable (Figure 1.5).
1.5 Conclusion

Using the external variable ‘score’ as a supplementary variable, the bootstrap resampling proves the stability of the relationship between the first principal coordinate vector and the wine score. The latter has been considered as a quantitative variable but also as a categorical variable, through grouping the values into categories.

The analysis of a lexical table through CA benefits from the validation of the structure by using the bootstrap procedure on both active and supplementary columns.

Software note
Bootstrap simulations as well as statistical computations have been carried out by means of specific software developed by the authors called SIMTEXT. This software runs under Windows and can be downloaded free from:

http://www3.unileon.es/personal/wwdderae/simtext/publish.htm
Supplementary Elements on Principal Axes Maps

Figure 1.5. Replicated correlations between the score and the first principal coordinate vector.

Acknowledgements. This work has been partially supported by the Spanish Ministry of Education and Science, FEDER (Grant SEJ2005-00741/ECON) as well as the Catalan Commission for the Universities DURSI (Grant SGR 00004/2005) and Junta of Castile and León (E-107/2004). We acknowledge the Wine Guide editor Mr. Fernando Lázaro for providing us information.

References


A Doubly Projected Analysis for Lexical Tables

Simona Balbi¹ and Michelangelo Misuraca²

¹ Dipartimento di Matematica e Statistica, Università di Napoli “Federico II,” Italy
² Dipartimento di Economia e Statistica, Università della Calabria, Italy

Abstract: This chapter aims to show how external information contributes in analysing a lexical table by enriching the readability of factorial maps. The theoretical frame is given by principal component analysis onto a reference subspace, a method based on the orthogonal projection of a correlation structure on the space spanned by an external set of explanatory variables. In previous papers the idea of a projected lexical analysis has been introduced by using a single reference space for terms. Here we consider a double projection strategy by involving external informative structures both on documents and terms, i.e., on rows and columns of a lexical table.

Keywords and phrases: External information, orthogonal projectors, factorial maps

2.1 Introduction

The necessity of introducing additional information in exploring multivariate structures, by means of Principal Component Analysis (PCA) and related techniques, has been a recurring topic in the literature since C. R. Rao introduced a set of instrumental (explanatory) numerical variables in PCA (Rao, 1964).

In textual data analysis this state is particularly pressing, due to the nature of the data. A preprocessing step is always necessary in analysing a document collection from a statistical standpoint. This process often allows one to reduce the linguistic variability in the data, considered in terms of noise for analysis purposes, but on the other hand it leads to a loss of information for comprehension of the phenomenon.

We often have information dealing with the document categorisation process and the context in which terms have been used. This external information (also known as metadata) can be used both in an informal way to aid subjective interpretations of the results and in a formal way by incorporating it in the data.

It is possible to consider two different kinds of information in a textual analysis:

• Intratextual information, usually quantitative and corpus-driven, that takes into account the relationships between terms and documents
• **Extratextual information**, usually qualitative, involving all those aspects strictly linked to the context in which the documents are produced not directly readable from the dataset.

The introduction of additional information both on individuals and variables, by considering two external informative matrices, was proposed by Takane and Shibayama (1991) combining features of regression analysis and PCA. The method was developed afterwards by Takane in the so-called Constrained Principal Component Analysis (CPCA) (Takane, 1997).

Sharing the data structure and having in mind CPCA properties and metrics considerations, in the following we choose as our methodological starting point another method, i.e., Principal Component Analysis onto a Reference Subspace (PCAR) (D’Ambra and Lauro, 1982), in order to emphasise the geometrical approach in terms of orthogonal projectors. After reviewing the main issues on the topic, in the following we propose a double projection strategy, simultaneously using orthogonal projectors on the spaces spanned by the additional variables related to documents and to terms. The effectiveness of the proposed strategy is shown by analysing the educational offerings of the Italian University.

### 2.2 Some methodological recall

In geometry an **orthogonal projection** of a $k$-dimensional object onto a $d$-dimensional subspace spanned by the $d$ columns linearly independent of a matrix $P(n,d)$, is obtained by considering a projection operator $P (P' P)^{-1} P'$, symmetric and idempotent. From a statistical viewpoint projecting a data structure onto a reference subspace means to analyse the relations between the rows and the columns in the frame of the information listed in $P$.

**2.2.1 Constrained principal component analysis**

CPCA data structure is given by an individual-by-variable matrix $Z$, and two external information matrices, $G$ (on individuals) and $H$ (on variables). According to Takane a wide variety of multivariate statistical analyses different from PCA are considered as interesting special cases of CPCA, including, e.g., correspondence analysis.

It has been thought of as a comprehensive method. For that reason there are no prescriptions in terms of distributional assumptions, preprocessing, or metric choices. The individual empirical interests suggest the proper behaviour to researchers. CPCA consists in two main analytical steps. In the first one, the so-called **external analysis**, $Z$ is orthogonally projected onto the spaces spanned by $G$ and $H$, in order to decompose the influence of the “external” variables into the sum of four terms: the first one pertains to what can be jointly explained by $G$ and $H$, the second one and the third one, respectively, pertain to what can be explained by $G$ and $H$, while the fourth one is a matrix of residuals. This solution is achieved in a least square estimation framework by minimising the residual matrix. In the second step the **internal analysis** is performed on the decomposition matrices by means of one or more PCA.

### 2.2.2 Principal component analysis onto a reference subspace

PCAR data structure is given by two individuals-by-variables matrices $Z$ and $X$. PCAR aims at visualising, in a proper geometrical framework, the dependence of $Z$ on $X$.

Namely PCAR looks for the principal components of the orthogonal projection of $Z$ on the space spanned by the columns of $X$. It can be seen as a special case of a CPCA internal analysis, when only the first term of the decomposition is considered and we want to introduce external information only on variables. Moreover the variables in $Z$ are centered and frequently standardised. In this sense it is a proper PCA. The advantages of PCAR are strictly connected to graphical aspects and interpretation. In fact factorial maps show both the correlations within the same set of variables and the correlations between the two sets.

### 2.3 Basic concepts and data structure

External information on both documents and terms can help in explaining the use of some keywords under defined conditions. We can focus our attention on the residual uses in order to enhance peculiarity in the terms used in single documents, not connected to the main interpretation keys.

Suppose we are considering two indicator matrices $I(n,j)$ and $J(n,j)$, representing two categorical variables observed on the same set of individuals. Let $N(I,j)$ be the contingency table cross-classifying the variables in $I$ and $J$. In the frame of a textual statistics viewpoint $N = I J$ is a lexical table having $I$ documents in rows and $J$ terms in columns. Correspondence Analysis (CA) is usually performed on this table (Lebart et al., 1997) in order to analyse and graphically represent the latent lexical relationships between documents and terms.

Frequently in analysing document collections we dispose of additional information concerning a possible categorisation scheme for documents and about the context in which terms are used. Let us consider therefore an indicator matrix $Y(I,K)$, assigning each document to a category $k (k = 1, \ldots, K)$. It is possible to perform a CA on the so-called aggregated lexical table $T(K,J)$ obtained as the dot product of $Y$ and $N$, in order better to read the relationships between groups of similar documents and the terms.

In previous papers, e.g., Balbi and Giordano (2000) and Balbi et al. (2002), the introduction of external information on documents and terms has already been performed for emphasising the different role played in the analysis. Here we focus mainly on an internal analysis in the sense of CPCA, stressing the geometrical features proper of PCAR. In other terms, by means of orthogonal projectors we want to visualise on factorial maps the association structure in $N$ due to the external information.

Additionally to the introduced matrices $Y$ (containing information on documents) and $N$ (the lexical table), let us consider a matrix $X(J,L)$ containing information on the vocabulary of the corpus we want to analyse. By using the residual part a context-independent representation is obtained. In Figure 2.1 the complete data structure is shown.
2.4 A doubly projected analysis

Let \( P_Y = Y(Y'Y)^{-1}Y' \) and \( P_X = X(X'X)^{-1}X' \), the orthogonal projectors considered both in CPC and in PCAR. Our proposal consists in analysing the residual matrix \( A = (I_X - P_Y)'N(I_X - P_X) \).

If we consider the matrices \( Y \) and \( J \) cross-tabulated in \( N \) it could be possible to split the matrix \( A \) in a first term \((X'X)^{-1}X'Y\) and in a second one given by \( YY'Y \), the profiles matrices containing the conditional distribution of \( Y \) on \( X \) and the conditional distribution of \( J \) on \( Y \), respectively. Those two matrices are the basic matrices of PCAR or more properly in its version developed for categorical variables, known as Non-Symmetrical Correspondence Analysis (NSCA) (Lauro and D'Ambra, 1984).

In other terms, we jointly study the residual part of dependence of the terms' distribution, with respect to the external information in \( Y \), and the residual part of dependence of the documents' distribution, with respect to the external information in \( X \). Dealing with elementary elements (i.e., each single occurrence in the corpus) it is worth noting that the dimensions of the two matrices are very huge, therefore this approach is infeasible in practice but would be useful in understanding results. In a similar way we can decide to carry out our analysis on any of the matrices considered by Takane for internal analysis.

The reference to PCAR is very useful in graphically visualising the results, because it makes it possible to represent on the factorial maps all the elements taken into account, the association structure in \( N \) together with the distribution of the terms conditioned to the information on \( Y \) and the distribution of documents conditioned to the information on \( X \).

2.5 The Italian academic programs: A study on skills and competences supply

In the frame of the European Union harmonisation policies the Italian university system has been reformed in 2000/2001 by introducing a new academic organisation. Two different kinds of degree have been introduced, a first-level three-year course (Laurea Triennale) followed by a specialising two-year course (Laurea Specialistica). All the courses are classified in 47 and 109 categories, respectively, divided among four general areas (humanities, social, scientific, and medical issues). The Ministry for Research and University has prepared for each course an explanatory declaration with the main contents and matters, used as a model by the University for planning their specific academic programs.

According to a request of the National Board for University System Evaluation, an official advisory body named by the ministry, the entire collection of academic programs has been analysed for investigating the correspondence between the competences and the skills supplied in the different courses and the employing outlets.

In this work we focus on the first-level academic programs. We have extracted 2812 declarations involving the different courses and in particular the fragments related to the competences offered have been selected. In this way the corpus of interest contains about 800,000 occurrences with a 17,200-term vocabulary. The lexical richness is not very wide, because in many cases the universities have decided to follow exactly the original declarations in drawing up the course programs.

The external information we introduce is the 47 ministerial course classes for the declarations (documents) and an eight-class competences categorisation obtained by previous analyses (terms).

Results of the PCA are strictly dependent on the choices of preprocessing procedures (centering, standardising, etc.) on the metrics used for computing distances and on the weighting systems. In this case we want to read the nontrivial uses of terms for describing competences in a context-independent framework, so that in the decomposition we assume an Euclidean metric and unitary weights. By introducing in the factorial coordinates a weighted Euclidean metric it is possible to recover the comparability in a common scale.

In Figure 2.2 we try to examine the peculiarities in the single courses' descriptions, independent of both the kind of competences offered and on the courses' nature. It is interesting to note how the first factorial axis (\( \approx 11\% \)) opposes general and abstract nouns (attività, competenze, formazione) on the right side, to terms describing more practical activities (predisposizione di progetti, ambiti differenziati, azioni di pianificazione, uso di tecnologie). This can be seen as a proper characterisation of the different university programs, in terms of a teaching frame oriented to "technical knowhow" or to a "way of thinking", which is a much-discussed topic in Italy. A deeper insight on this question is given by the second axis (\( \approx 7\% \)), where we can see an opposition, from the top to the bottom, between basic and professional/technical competences.

The interpretation of this map is coherent with the European debate about the new role of the university, in which it is necessary to include in the academic programs not only theoretic knowledge (to know) but more and more technical skills (to know how). It can be possible to perform the same analysis also on the specialising courses and to compare the two configurations of points in order to evaluate the internal coherence of the formative projects offered by the universities in terms of general and specific competences.
Figure 2.2. Factorial representation on the first two axes (≈18% of explained inertia)

References


