Multiple Correspondence Analysis for the Quantification and Visualization of Large Categorical Data Sets

Article · December 2012

DOI: 10.1007/978-3-642-21037-2_41

3 authors, including:

Alfonso Iodice D'Enza
University of Naples Federico II
32 PUBLICATIONS **432** CITATIONS

Michael Greenacre
University Pompeu Fabra
180 PUBLICATIONS **12,000** CITATIONS

Multiple Correspondence Analysis for the Quantification and Visualization of Large Categorical Data Sets

Alfonso Iodice D'Enza and Michael Greenacre

Abstract The applicability of a dimension-reduction technique on very large categorical data sets or on categorical data streams is limited due to the required singular value decomposition (SVD) of properly transformed data. The application of SVD to large and high-dimensional data is unfeasible because of the very large computational time and because it requires the whole data to be stored in memory (no data flows can be analysed). The aim of the present paper is to integrate an incremental SVD procedure in a multiple correspondence analysis (MCA)-like procedure in order to obtain a dimensionality reduction technique feasible for the application on very large categorical data or even on categorical data streams.

1 Introduction

In social, behavioural, environmental sciences as well as in marketing, a large amount of information is gathered and coded in several attributes. In most cases the aim is to identify pattern of associations among the attribute levels. A data-mart selected from a categorical data base can be represented as a binary data matrix whose rows correspond to records and whose columns correspond to the levels of the attributes.

A well-known exploratory method to describe and visualize this type of data is multiple correspondence analysis (MCA) (Greenacre 2007). MCA is widely

A.I. D'Enza (✉)
Università di Cassino, Cassino, Italy
e-mail: iodicede@gmail.com

M. Greenacre
Universitat Pompeu Fabra, Barcelona, Spain
e-mail: michael@upf.es

A. Di Ciaccio et al. (eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-21037-2_41,

applied in different fields: from marketing to social sciences, behavioural and environmental sciences. MCA is the generalization of correspondence analysis (CA) to more than two categorical variables. CA and MCA can be viewed as an adaptation to categorical data of principal component analysis (PCA, Jolliffe (2002)). As PCA, MCA aims to identify a reduced set of synthetic dimensions maximizing the explained variability of the categorical data set in question. The advantages in using MCA to study associations of categorical data are then to obtain a simplified representation of the multiple associations characterizing attributes as well as to remove noise and redundancies in data. The exploratory and visualization-based approach characterizing MCA provides immediate interpretation of the results.

Roughly speaking the MCA implementation consists of a singular value decomposition (SVD) – or the related eigenvalue decomposition (EVD) – of properly transformed data. The applicability of MCA on very large data sets or on categorical data streams is limited due to the required SVD. The application of the SVD to large and high-dimensional data is unfeasible since it requires a computational time that is quadratic in the data size; furthermore the SVD input matrix must be complete (no missings) and stored in memory (no data flows can be analysed) (Brand 2003). Since the SVD characterizes many techniques aiming at dimension reduction, noise suppression and clustering, many contributions in the literature aim to overcome the SVD computational problem by updating the SVD incrementally. For example Zhao et al. (2006) propose a scalable dimension-reduction technique for continuous data such as incremental PCA that exploits SVD updating procedures.

In some applications the data set to be analysed can be stratified in different subgroups. The need for splitting data in chunks is two-fold: (a) if the amount of data to analyse is very large, or data are produced at a high rate (data flows), it can be convenient or necessary to process it in different "pieces"; (b) if the data in question refer to different occasions or positions in time or space, a comparative analysis of data stratified in chunks can be suitable.

The aim of the present contribution is to propose an MCA-like procedure that can be updated incrementally as new chunks of data are processed. The proposed procedure is obtained by integrating an incremental SVD with a properly modified MCA procedure. The low-dimensional quantification and visualization of categorical attributes via this MCA-like procedure is a promising approach to investigate the association structures and for fast clustering purposes.

The present paper is structured as follows: in Sect. 2 the motivation of stratifying data into chunks is described; Sect. 2.1 briefly recalls the computations of MCA and in Sect. 2.2 some options are recalled to study the association structures along the different chunks. Section 3 contains a description of the basic procedure for incremental SVD. Section 4 described the modification to the MCA procedure necessary to embed the incremental SVD procedure. An application is proposed in Sect. 5 and a last section is dedicated to future work.

2 Multiple Correspondence Analysis of Data Chunks

The stratification of a data set into different chunks can be performed for computational reasons, when the data set is too large to be analysed as a whole. Depending on the context, chunks can be defined according to an external criterion, related to time or another characteristic. For example, consider the market basket analysis framework. Since the aim is to study the buying behaviours of customers, then a suitable task is to monitor customer choices along weeks or months, as well as to appreciate customer reactions to promotion policies. Chunks are determined in this case according to time or to different promotions on products.

Furthermore, in evaluating higher-education systems, CA or MCA can be suitably applied to analyse student careers (with attribute levels indicating whether or not an examination is passed): the stratification in this case could be used to compare the overall behaviour of students from different universities or at different academic years. Further applications involving large amounts of categorical data are in text mining and web-clickstream analysis among others.

2.1 *Computations of Multiple Correspondence Analysis*

Let us consider a data-mart resulting from the selection of n records and Q attributes from a categorical data base. Let $J_q, q = 1, \ldots, Q$, be the levels of each categorical attribute. The indicator matrix $\mathbf{Z}$ has n rows and J columns, where $J = \sum_{q=1}^{Q} J_q$. The general element is $z_{ij} = 1$ if the ith record assumes the jth attribute level, $z_{ij} = 0$ otherwise.

Two possible approaches to MCA consist of the application of CA algorithm to the indicator matrix $\mathbf{Z}$ and to the Burt matrix $\mathbf{C} = \mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ (see appendix of Greenacre and Blasius (2006) for details). Let us briefly recall the computations based on the Burt matrix: let the correspondence matrix be $\mathbf{P} = \frac{\mathbf{C}}{(n \times Q^2)}$, with $(n \times Q^2)$ being the grand total of $\mathbf{C}$. Let the vector $\mathbf{r}$ contain the row sums of $\mathbf{P}$, which are also column sums since $\mathbf{P}$ is symmetric. In MCA, as well as in CA, the association structure is revealed by performing a SVD (or equivalently EVD, in this particular case) of the standardized residuals matrix

$$\mathbf{S} = \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{r}^{\mathsf{T}} \right) \mathbf{D}_r^{-1/2} \tag{1}$$

where $\mathbf{D}_r$ is the diagonal matrix of the elements of $\mathbf{r}$. The general element of the $\mathbf{S}$ matrix is then

$$s_{ij} = \frac{(p_{ij} - r_i r_j)}{\sqrt{r_i r_j}}. \tag{2}$$

The SVD of $\mathbf{S}$ is

$$\mathbf{S} = \mathbf{U}\Sigma\mathbf{U}^{\mathsf{T}} \tag{3}$$

where Σ is the diagonal matrix of singular values in descending order, and $\mathbf{U}$ is the matrix of singular vectors, respectively. Note that only the $J - Q$ positive singular values are retained. The reduced space representation of row and column profiles is obtained via the SVD results. In particular, the principal coordinates of the rows, or columns since $\mathbf{S}$ is symmetric, are

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\Sigma. \tag{4}$$

2.2 Evolutionary MCA Solutions

The exploratory study of association of multiple categorical data chunks can be approached in different ways. The most basic and straightforward way is to perform a MCA on each data chunk. In this case, however, each MCA solution refers to a single chunk and a comparison among the solutions is difficult, since the displays refer to different subspaces.

An important option is the three-way correspondence analysis (Carlier and Kroonenberg 1996) for the analysis of three-way contingency tables. Although this approach is particularly useful in studying association structures at different occasions, chunks are supposed to have same size in order to consider cubes of data. A further approach exploiting some of the properties of CA (and MCA) aims at a comparative study of association structures. Once a chunk (or a set of chunks) is selected as a reference analysis, further chunks are projected as supplementary information, as proposed by Iodice D'Enza and Palumbo (2007, 2009) and Palumbo and Iodice D'Enza (2009). This approach allows a direct comparison among the displays of the chunks since they all refer to the same reference subspace. The weakness of this approach is in the prominence of the associations characterizing the reference chunks: the resulting subspace is constant and it conditions all the further visualizations. Then a quite natural step forward is to update the reference MCA solution incrementally as new chunks are processed in order to have a comparable and updated subspace of projection.

3 Enhanced Methods for SVD

The SVD is computed via a batch time algorithm with a computational complexity $O\left(nQ^2 + n^2Q + Q^3\right)$, (Golub and van Loan 1996): then SVD becomes unfeasible as the values of n and Q increase. Furthermore, because all of available data are necessary to obtain the decomposition, it cannot be suitably applied to data flows.

The contribution can be roughly categorized in batch methods and incremental methods. The batch methods aim to reduce the computational effort of the SVD extending its applicability: the Lanczos bilinear diagonalization methods (Baglama and Reichel 2007) represent a relevant example of such a contribution.

The incremental approach aims to update an existing SVD solution when new data comes in. These methods have the advantage over the batch methods as they can be applied to subsequent chunks of data without the need to store the previously analysed data in memory. This has motivated the description of numerous SVD updating methods, e.g., the contribution by Businger (1970) and by Bunch and Nielsen (1978). The SVD update process, although of high complexity, presents several advantages. Most updating procedures rely on the *dominant* SVD which is a decomposition retaining the r largest singular values and the related singular vectors: this is done in order to reduce the overall computational complexity. Examples of such procedures are in the proposals by Chandrasekaran et al. (1997), Levy and Lindenbaum (1998), Chahlaoui et al. (2001), Kwok and Zhao (2003) and Brand (2003, 2006). These methods approximate the dominant SVD after a single scan of the matrix, and they maintain only a low-rank factorization: that is, these methods are able to approximate the SVD using less memory and computation than direct full-rank SVD methods (Baker et al. 2008).The reference procedure is the online SVD proposed by Brand (2003; 2006) for updating the decomposition when additive modifications occur on the starting data matrix: this kind of modification is particularly useful in embedding the SVD updates in a MCA procedure. The main steps of the procedure are described in Sect. 3.1.

3.1 Incremental SVD Procedure

In the present paper we refer to the online SVD procedure proposed by Brand (2003) in the context of recommender systems. Consider a $n \times p$ continuous data matrix $\mathbf{X}$ and its *rank-r* SVD ($\mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}$) representing an approximation of $\mathbf{X}$ according to the r highest singular values. Let $\mathbf{A}$ and $\mathbf{B}$ be two modification matrices with c columns each and n and p rows, respectively. The aim is to obtain the SVD of $\mathbf{X} + \mathbf{A}\mathbf{B}^{\mathsf{T}}$ by updating $\mathbf{U}$, Σ and $\mathbf{V}$. The detailed explanation of the procedure is in the contributions by Brand, in the appendix of Brand (2003) and in Brand (2006). Here follow the most relevant steps:

(a) Perform a modified Gram-Schmidt procedure for the QR decomposition as follows:

$$[\mathbf{U}\;\mathbf{A}] \overset{QR}{\rightarrow} [\mathbf{U}\;\mathbf{P_A}] \begin{bmatrix} \mathbf{I} & \mathbf{U}^{\mathsf{T}}\mathbf{A} \\ \mathbf{0} & \mathbf{R}_A \end{bmatrix},$$

$$[\mathbf{V}\;\mathbf{B}] \overset{QR}{\rightarrow} [\mathbf{V}\;\mathbf{P_B}] \begin{bmatrix} \mathbf{I} & \mathbf{V}^{\mathsf{T}}\mathbf{B} \\ \mathbf{0} & \mathbf{R}_B \end{bmatrix}; \tag{5}$$

note that the QR decomposition factorize a matrix into an orthogonal matrix and an upper triangular matrix. The matrices $\mathbf{P}_A, \mathbf{Q}_A, \mathbf{P}_B, \mathbf{Q}_B$ are blocks of the matrix resulting from the QR decompositions.

(b) Consider the following matrix

$$\mathbf{K} = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{U}^{\mathsf{T}}\mathbf{A} \\ \mathbf{R}_A \end{bmatrix} \begin{bmatrix} \mathbf{V}^{\mathsf{T}}\mathbf{B} \\ \mathbf{R}_B \end{bmatrix}^{\mathsf{T}} \tag{6}$$

and compute the *rank*-$(r + c)$ SVD of $\mathbf{K}$

$$\mathbf{K} = \mathbf{U}^{'} \Sigma^{'} \mathbf{V}^{'\mathsf{T}}. \tag{7}$$

(c) The SVD of $\mathbf{X} + \mathbf{A}\mathbf{B}^{\mathsf{T}}$ is obtained as

$$\mathbf{X} + \mathbf{A}\mathbf{B}^{\mathsf{T}} = \left([\mathbf{U}\ \mathbf{P}_{\mathbf{A}}]\, \mathbf{U}^{'} \right) \Sigma^{'} \left([\mathbf{V}\ \mathbf{P}_{\mathbf{A}}]\, \mathbf{V}^{'} \right)^{\mathsf{T}} = \mathbf{U}^{''} \Sigma^{'} \mathbf{V}^{''\mathsf{T}}. \tag{8}$$

The matrix $\mathbf{X}$ has the same singular values as $\mathbf{K}$ whereas the left and right singular vectors depend on both the corresponding singular vectors of the matrix $\mathbf{X}$ ($\mathbf{U}$ and $\mathbf{V}$) and those of the matrix $\mathbf{K}$ ($\mathbf{U}'$ and $\mathbf{V}'$). Note that the procedure requires a rank-$(r + c)$ SVD of $\mathbf{K}$ instead of a rank-r SVD of a pre-updated matrix $\mathbf{X}$. However, as pointed out by Brand (2006), the matrix $\mathbf{K}$ is sparse and highly structured, so it is easily diagonalized. Furthermore, the SVD update does not require the starting matrix $\mathbf{X}$ to be kept in memory, only the starting SVD and $\mathbf{A}$ and $\mathbf{B}$ matrices are needed.

4 Update of MCA-Like Results

In order to integrate the SVD update procedure in MCA it is necessary to define both $\mathbf{A}$ and $\mathbf{B}$ matrices to reflect new upcoming data. In particular, let the starting matrix be $\mathbf{X} = \mathbf{S}$, the original matrix of standardized residuals. $\mathbf{A}$ and $\mathbf{B}$ must be such that $\mathbf{A}\mathbf{B}^{\mathsf{T}}$ contains the additive modifications of $\mathbf{S}$ when new rows of $\mathbf{Z}$ are added. Let $\mathbf{Z}^{+}$ be the indicator matrix of upcoming data and let $\mathbf{C}^{+} = \mathbf{Z}^{+\mathsf{T}}\mathbf{Z}^{+}$ the corresponding Burt matrix. Since

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^{+} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^{+} \end{bmatrix} = \mathbf{C} + \mathbf{C}^{+}$$

the aim is to properly transform $\mathbf{C}^{+}$ to update the starting matrix $\mathbf{S}$. Let n^{+} be the number of added rows, with $\left[(n + n^{+}) \times Q^2\right]$ being the updated *grand total*. Define the correspondence matrix update as

$$\mathbf{P}^{+} = \frac{\mathbf{C}^{+}}{(n + n^{+}) \times Q^{2}}. \tag{9}$$

Let $\mathbf{r}^{+}$ be the vector of rows (columns) margins of $\mathbf{P}^{+}$. The standardized residuals updating matrix is

$$\mathbf{S}^{+} = \mathbf{D}_{r}^{-1/2} \left(\mathbf{P}^{+} - \mathbf{r}^{+}\mathbf{r}^{+\mathsf{T}} \right) \mathbf{D}_{r}^{-1/2}. \tag{10}$$

Note that the centring operators are the update of the margins, while the residuals are divided by the original independence condition: this is done in order to keep the updating quantity in the same scale as the original matrix $\mathbf{S}$. This update does not reconstruct exactly the standardized residual matrix. Then the modification matrices to input in the SVD update are defined as

$$\mathbf{A} = \mathbf{D}_{r}^{-1/2} \left(\mathbf{P}^{+} - \mathbf{r}^{+}\mathbf{r}^{+\mathsf{T}} \right) \text{ and } \mathbf{B} = \mathbf{D}_{r}^{-1/2}.$$

The actual residuals are divided by the starting independent condition instead of the updated one, this leads to some difference in the usual interpretation of the attribute points on the map. In particular, the centring operator is updated, then the centre of the map still represents the independence condition. The distance of the points from the centre is still a weighed Euclidean metric, but it is not a chi-square distance, since the weights are not updated. However, this affects the scale of the configuration of points but not the relative position of the points.

5 Example of Application

The data set is taken from the multinational ISSP survey on environment in 1993 and it is provided with the R package ca (Nenadić and Greenacre 2007). The number of considered attributes is $Q = 7$, the number of respondents is $n = 871$. In particular, attributes correspond to four substantive questions from the ISSP survey (see Table 1), then there are three demographic variables such as gender, age and education. The total number of attribute levels is $J = 34$. The possible answers to each question are: (1) strongly agree, (2) somewhat agree, (3) neither agree nor disagree, (4) somewhat disagree, (5) strongly disagree (notice that it is usual practice to not mix substantive and demographic variables as active variables, but we do so here merely to illustrate the technical functioning of our procedure). Although the data dimensionality is quite small, the aim is to show the procedure on real data and compare the results with those of a regular MCA on the whole data matrix. In order to update incrementally a starting solution we randomly selected $n_0 = 100$ respondents. The remaining 771 respondents were treated as further information and split into $k = 10$ chunks of approximately equal sizes.

Table 1 The ISSP survey data set

Attribute	n. of levels
A: We believe too often in science, and not enough in feelings and faith	5
B: Overall, modern science does more harm than good	5
C: Any change humans cause in nature, no matter how scientific, is likely to make things worse	5
D: Modern science will solve our environmental problems with little change to our way of life	5
Gender	2
Age	6
Education	6

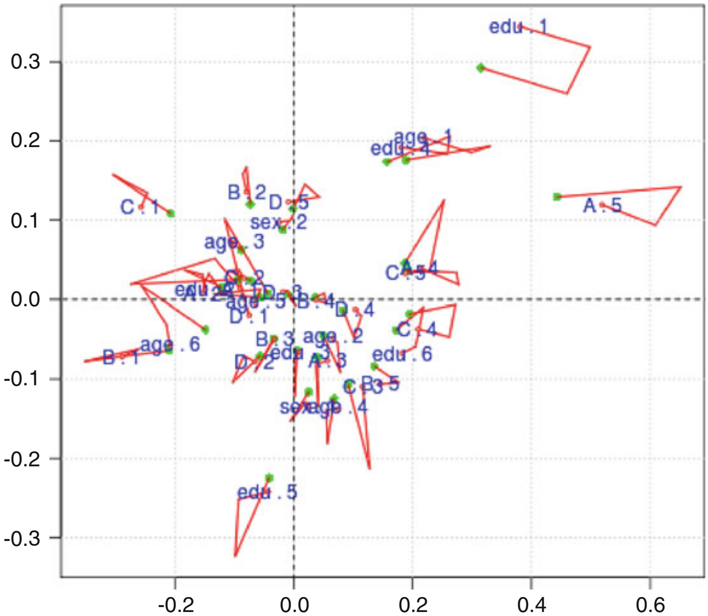

Fig. 1 Trajectories of the attribute points: starting, two intermediates and final positions

Figure 1 shows the trajectories of the attribute points: each broken line goes from the starting position (corresponding to the first n_0 respondents), then passes through the third and the sixth intermediate frames and ends at the final position of the corresponding attribute level. The attribute labels are placed at the end of the corresponding trajectory. What Fig. 1 shows is that the amount of change in the attribute level positions is proportional to their distance from the centre of the map. Since the centre of the map corresponds to the independence condition, attributes placed close to the centre of the map are characterized by a lower level of association with one other. Then highly associated attribute levels are more "sensitive" to the new data

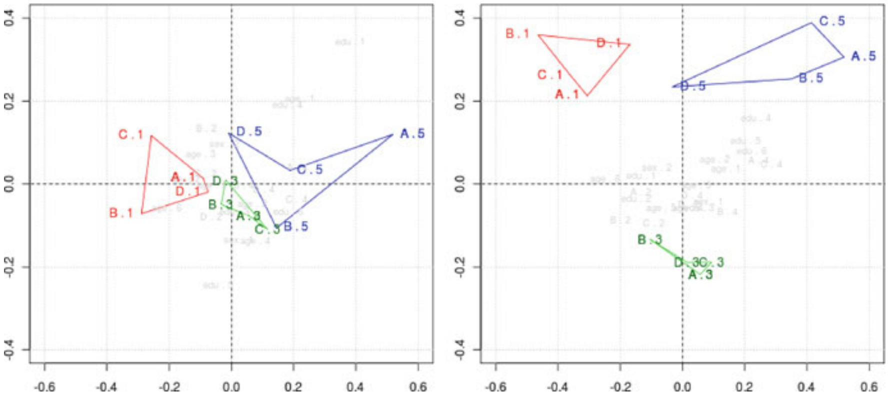

Fig. 2 Solution of the incremental MCA-like procedure (*left*) and of the MCA (*right*)

being incorporated as their level of association is more likely to change. The evolutionary structure of association could be better detected by a dynamic visualization, showing the frame-by-frame configuration of points. As the trajectory shows, there are some changes from one frame to another, but the overall structure of association of the whole data set does not change in a relevant way. This is due to the fact that there is a well-defined association structure underlying the data set in question.

Some further comments arise when comparing the incrementally updated solution with the MCA solution resulting for the analysis of the whole data set. In order to ease the comparison, as Fig. 2 shows, we linked together the points corresponding to the response "strongly agree" for each of the four substantive questions. We did the same for the responses "neither agree nor disagree" and "strongly disagree". Looking at the three resulting polygons in the left- and right-hand side of Fig. 2 it can be seen that the relative positions of the polygons are fairly similar and they present the "horse shoe" effect which is typical of CA and MCA. Although the comparability of the obtained result with MCA is a desirable feature, it has to be noted that, in the case of the incremental MCA-like solution, the polygons are less spread over the map compared to the MCA map. This is due to the update of the Burt matrix (see formula (10)): in particular, for sake of comparability, the updating residuals are standardized by the starting marginals, and this condition keeps the axis scale artificially fixed. This is not a desirable feature, so more enhanced updates are needed.

6 Future Work

The proposed approach is mainly aimed to extend the usability of a powerful exploratory and visualization-based technique like MCA to a wider range of application. Future developments are needed and further enhancements are possible, for example to enhance reconstruction of the updated standardized residual matrix. Another development of the procedure will be to focus on the computations based on the indicator matrix rather than the Burt matrix: this makes it possible to use lower rank modifications of the SVD that require a lower computational effort. In order to make more comparable the frame-by-frame visualizations the idea is to adopt a Procrustean rotation matrix that compensates for the "rotation effect" due to the subspace update. One last step is to exploit the SVD update to take missing values into account.

References

J. Baglama and L. Reichel (2007). Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM J. Sci. Comput.*, 27, 19–42.

Baker C., Gallivan K. and Van Dooren P. (2008) Low-rank incremental methods for computing dominant singular subspaces. Computer Science and Mathematics Oak Ridge National Laboratory (**url:** http://www.csm.ornl.gov/~cbaker/Publi/IncSVD/copper08.pdf).

Brand M. (2003). Fast online svd revision for lightweigth recommender systems. *SIAM International Conference on Data Mining*.

Brand M. (2006). Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra and its Applications*, 415, 20–30.

Bunch J. R. and Nielsen C. P. (1978). Updating the singular value decomposition. *Numerische Mathematik*, 31(2):111–129.

Businger P (1970). Updating a singular value decomposition. *BIT*, 10(3), 376–385.

Carlier A. and Kroonenberg P.M. (1996). Decompositions and biplots in threeway correspondence analysis. *Psychometrika*, 61, 355–373.

Chahlaoui Y., Gallivan K. and Van Dooren P. (2001). An incremental method for computing dominant singular spaces. *Computational Information Retrieval*, SIAM., 53–62.

Chandrasekaran S., Manjunth B. S., Wang Y.F., Winkeler J. and Zhang H. (1997). An eigenspace update algorithm for image analysis, *Graphical Models and Image Processing*, 59(5), 321–332.

Golub G. and van Loan A. (1996). *Matrix Computations*, John Hopkins U. Press.

Greenacre M.J. (2007). *Correspondence Analysis in Practice*, second edition, Chapman and Hall/CRC.

Greenacre M.J. and Blasius J. (2006). *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC, first edition.

Jolliffe I.T. (2002). *Principal Component Analysis*, Springer-Verlag, second edition.

Kwok J. and Zhao H. (2003). Incremental eigen decomposition. *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 270–273, Istanbul, Turkey.

Iodice D'Enza A. and Palumbo F. (2009). A regulatory impact analysis (RIA) approach based on evolutionary association patterns. *Statistica Applicata*, accepted, in press.

Iodice D'Enza A. and Palumbo F. (2007). Binary data flow visualization on factorial maps. *Revue Modulad*, n.36.

Levy A. and Lindenbaum, M. (1998). Sequential Karhunen-Loeve basis extraction and its application to images. *Technical Report CIS9809*, Technion.

Nenadić O. and Greenacre M.J. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3), 1–13.

Palumbo F. and Iodice D'Enza A. (2009). Clustering and dimensionality reduction to discover interesting patterns in binary data. *Advances in Data Analysis, Data Handling and Business Intelligence*. Springer, 45–55.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Zhao H., Chi P. and Kwok J. (2006). A novel incremental principal component analysis and its application for face recognition. *Systems, Man and Cybernetics, Part B: Cybernetics, IEEE Transactions*, 35, 873–886.