

Multiple Correspondence Analysis for the quantification and visualization of huge categorical data

Alfonso Iodice D'Enza¹

Dipartimento di Scienze Economiche, Università di Cassino, e-mail: iodicede@unina.it

Michael J. Greenacre

Departament d'Economia i Empresa, Universitat Pompeu Fabra, e-mail: michael@upf.es

Abstract: The applicability of a dimensionality 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 analyzed). The aim of the present paper is to integrate the incremental SVD procedure proposed by Brand (2003) 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.

Keywords: Categorical data, Correspondence Analysis, Dimensionality Reduction

1. Introduction

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 categorical attributes. A well known exploratory method to describe and visualize such kind of data matrix is Multiple Correspondence Analysis (MCA) (Greenacre, 2007). MCA is widely applied in social, behavioral and environmental sciences. MCA analysis 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). As PCA, MCA aims to identify a reduced set of synthetic dimensions maximizing the explained variability.

The advantages in using MCA to study associations of categorical data are then to obtain a synthetic representation of multiple associations characterizing attributes as well as to remove noise and redundancies in data.

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

¹Address of correspondence: via S. Angelo a Folcara, Cassino.

SVD updating procedures.

The aim of the present paper is to integrate the incremental SVD procedure proposed by Brand (2003) in a 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. The low dimensional quantification and visualization of categorical attributes via MCA-like is a promising approach to investigate the association structures and for fast clustering purposes. The present paper is structured as follows: in section (1.) the framework and the general aim of the paper is introduced; section (2.) briefly recalls MCA; the section (3.) contains a description of the incremental SVD basic procedure, the last section (4.) shows the SVD update integration in a MCA-like procedure and an example of application.

2. Computations of Multiple Correspondence Analysis

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

Two possible approaches to MCA consist of the application of CA algorithm to the indicator matrix Z and to the Burt matrix $C = Z^T Z$ (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 $P = \frac{C}{(n \times Q^2)}$, with $(n \times Q^2)$ being the *grand total* of C . Let the vector r contain the rows (columns) margins of P . In MCA, as well as in CA, the association structure is revealed by performing an SVD (or equivalently an EVD) of the standardized residuals matrix

$$S = D_r^{-1/2}(P - rr^T)D_r^{-1/2} \quad (1)$$

with D_r being diagonal matrix with the same elements as r . The decomposition of S leads to obtain

$$S = U\Sigma V^T \quad (2)$$

with Σ being the diagonal matrix of eigenvalues ordered in descending order: U and V being the left and right singular vectors, respectively. Remark that only the $J - Q$ positive singular values are retained. The reduced space representation of rows and columns profiles are obtained via the SVD results. In particular, the principal coordinates of rows (columns) are $F = D_r^{-1/2}U\Sigma$.

3. SVD update

In case of very large data set performing of SVD is unfeasible as it requires a considerable computational time for finite data sets and it is inapplicable to data flows. In the specific literature several contributions aim to update an SVD: 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 ($\mathbf{U}\Sigma\mathbf{V}^T$) representing an approximation of $\mathbf{X}$ according to the r highest eigenvalues. Let $\mathbf{A}$ and $\mathbf{B}$ two modification matrices with c columns each and n and p rows, respectively. The aim is to obtain the SVD of $\mathbf{X} + \mathbf{A}\mathbf{B}^T$ by updating $\mathbf{U}, \Sigma$ and $\mathbf{V}$. The detailed explanation of the procedure is in appendix of Brand (2003) and in Brand (2006) contributions. Here follow the most relevant steps:

i) perform a modified Gram-Schmidt procedure for the QR decomposition as follows:

$$[\mathbf{U} \ \mathbf{A}] \xrightarrow{QR} [\mathbf{U} \ \mathbf{P}_A] \begin{bmatrix} \mathbf{I} & \mathbf{U}^T \mathbf{A} \\ \mathbf{0} & \mathbf{R}_A \end{bmatrix} \text{ and } [\mathbf{V} \ \mathbf{B}] \xrightarrow{QR} [\mathbf{V} \ \mathbf{P}_B] \begin{bmatrix} \mathbf{I} & \mathbf{V}^T \mathbf{B} \\ \mathbf{0} & \mathbf{R}_B \end{bmatrix};$$

ii) consider the following matrix

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

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

$$\mathbf{K} = \mathbf{U}' \Sigma' \mathbf{V}'^T; \quad (3)$$

iii) the SVD of $\mathbf{X} + \mathbf{A}\mathbf{B}^T$ is obtained as

$$\mathbf{X} + \mathbf{A}\mathbf{B}^T = ([\mathbf{U} \ \mathbf{P}_A] \mathbf{U}') \Sigma' ([\mathbf{V} \ \mathbf{P}_B] \mathbf{V}')^T = \mathbf{U}'' \Sigma' \mathbf{V}''^T. \quad (4)$$

Remark that the SVD update does not require the starting matrix $\mathbf{X}$, 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 takes to properly 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 $\mathbf{A}\mathbf{B}^T$ contains the additive modifications of $\mathbf{S}$ when new rows of $\mathbf{Z}$ comes in. Let $\mathbf{Z}^+$ be the indicator matrix of upcoming data and let $\mathbf{C}^+ = \mathbf{Z}^{+T} \mathbf{Z}^+$ the corresponding Burt matrix. Since it results that

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

the aim is to properly transform $\mathbf{C}^+$ to update the starting matrix $\mathbf{S}$. Let n^+ be the number of added rows, with $[(n + n^+) \times Q^2]$ being the updated *grand total*. Define the correspondence matrix update as $\mathbf{P}^+ = \frac{\mathbf{C}^+}{(n+n^+) \times Q^2}$. Let $\mathbf{r}^+$ the vector of rows (columns) margins of $\mathbf{P}^+$. The standardized residuals updating matrix is

$$\mathbf{S}^+ = \mathbf{D}_r^{-1/2} (\mathbf{P}^+ - \mathbf{r}^+ \mathbf{r}^{+T}) \mathbf{D}_r^{-1/2}. \quad (5)$$

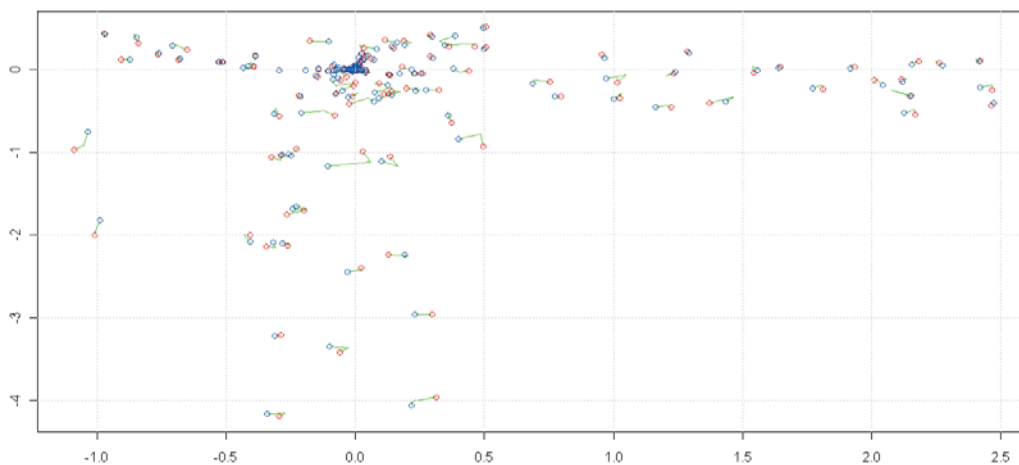
Notice that the centering operators are the update of the margins, while the residuals are divided by the original independence condition: this is done in order to keep the updating

quantity in the same scale as the original matrix S . Then the modification matrices to input in the SVD update are defined as

$$\mathbf{A} = \mathbf{D}_r^{-1/2}(\mathbf{P}^+ - \mathbf{r}^+ \mathbf{r}^{+\top}) \text{ and } \mathbf{B} = \mathbf{D}_r^{-1/2}.$$

In order to show an example of application of the procedure, we considered a well-known web-log data set available in sparse format on the web : data refer to the pages of Microsoft web-site visited by (anonymous) users in a week of February 1998. The starting analysis is on 10000 records of 87 two-levels attributes, then the quantification/visualization has been updated twice, adding each time 1000 records. The figure

Figure 1: Visualization of attributes at start (blue) and at the third update (red)



shows the attribute levels position at the start and after the second update, while the trajectory puts in evidence the intermediate update.

References

- Brand M. (2003) Fast online svd revision for lightweight recommender systems., in: *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.
- Greenacre M.J. (2007) *Correspondence Analysis in Practice, second edition*, Chapman and Hall/CR.
- Greenacre M.J. and Blasius J. (2006) *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC, first edition.
- 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 on*, 35, 873–886.