*Folia Oeconomica Stetinensia*
Volume 18 (2018) Issue 2
DOI: 10.2478/foli-2018-0026

WYDZIAŁ NAUK EKONOMICZNYCH
I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

# SINGULAR VALUE DECOMPOSITION APPROACHES IN A CORRESPONDENCE ANALYSIS WITH THE USE OF R

Justyna Brzezińska, Ph.D.

*University of Economics in Katowice*
*Faculty of Finance and Insurance*
*Department of Economic and Financial Analysis*
*1 Maja 50, 40-287 Katowice, Poland*
*e-mail: justyna.brzezinska@ue.katowice.pl*
*ORCID: https://orcid.org/0000-0002-1311-1020*

**Received 6 November 2017, Accepted 5 November 2018**

## Abstract

The aim of a correspondence analysis is the graphical representation of the categories of variables in one frame of reference. This visualization is possible due to the decomposition of the basic matrix with the use of Singular Value Decomposition (SVD). There are three matrices used in the process of decomposition: right singular vectors, left singular vectors, and a singular value diagonal matrix. The aim of this paper is to compare four different approaches and algorithms of SVD methods used in a correspondence analysis. In the literature, four approaches are known to singular value decomposition, defined by: R.A. Fisher (1940), M.J. Greenacre (1984), E.B. Anderson (1991), and J.D. Jobson (1992). Those computational procedures will be presented and compared in this paper. Also, methods of determining the coordinates of the category column and line matrix, as well as the values of inertia will be defined for these approaches. A key problem is to compare the well-known approaches, since in the literature only one approach – proposed by Greenacre – is used for singular value decomposition. The reason of the superiority of this algorithm over the others may be the simplicity and ease of the mathematical calculations. Greenacre's algorithm is also used in *R* statistical software, making its availability and popularity growing, however, other algorithms are worth presenting and focusing on.

**Keywords:** correspondence analysis, singular value decomposition, contingency table, R software

**JEL classification:** C35

## Introduction

A correspondence analysis is a multivariate statistical method designed for categorical variables in a contingency table. A correspondence analysis transforms a data table into two sets of new variables called factor scores (obtained as linear combinations of the rows and columns, respectively) – one set for the rows and one set for the columns. These factor scores give the best representation of the similarity structure of the rows and the columns of the table, respectively. In addition, the factor scores can be plotted as maps that optimally display the information in the original table. In these maps, rows and columns are represented as points whose coordinates are the factor scores, and where the dimensions are also called factors, components (by analogy with PCA), or simply dimensions. Interestingly, the factor scores of the rows and the columns have the same variance and, therefore, the rows and columns can be conveniently represented in one single map.

A graphical presentation of categories of variables is possible due to the decomposition of **A** matrix represented as the product of three matrices. This decomposition in three matrices allows to determine the coordinates of the categories, which enable a graphical presentation of results, as well as the value of inertia, which is a measure of the dispersion of points. There are four known algorithms of SVD of **A** matrix known in the literature: R.A. Fisher (1940), M.J. Greenacre (1984), E.B. Anderson (1991), and J.D. Jobson (1992). The aim of this work is to systematize and compare the results obtained through the use of the four presented algorithms. World literature is quite poor when it comes to systematizing the knowledge of SVD algorithm used in a correspondence analysis. There are also no comparisons between existing algorithms.

In this paper, we compare algorithms and basic concepts of a correspondence analysis, as well as SVD with graphical presentation. Also, the comparison of the four methods of SVD and the results obtained in each algorithm is described. We also present and compare graphical configuration points in two-dimensional space based on real-life data. All calculations are conducted in *R* software.

The study presented in this paper focuses on different SVD approaches and compares four algorithms. Due to the development of modern statistical software, it is necessary to learn more about Greenacre’s approach that is available in *R* software for a correspondence analysis. Since the end of the XX$^{\text{th}}$ century, there have been no new approaches nor computer procedures, that’s why the topic deserves to be examined. Whenever we will be forced to conduct SVD algorithm on our own, without automatic computer calculations, basic steps will be explained as well as graphical results using *R* will be given.

From all the modern approaches available nowadays in *R* only SVD (Korobeynikov, Larsen, 2016) and Rspectra (Qiu, Mei, Guennebaud, Niesen, 2016) packages provide competitive functions for computing the partial SVD in *R* software. Also, the SVD analysis and discussion can be found in G.W. Stewart (1993), M.J. Greenacre (2010), and S. Voronin, P.G. Martinsson (2015).

## 1. Correspondence analysis

Let **N** denote $I \times J$ data matrix, with positive row and column sums to N, consisting of nonnegative numbers. In the correspondence analysis, a correspondence matrix **P** will be used, which refers to the probability matrix $\mathbf{P} = \frac{1}{n}\mathbf{N}$. Row masses are defined as: $r_i = \sum_{j=1}^{J} p_{ij}$ , and column masses are defined as $c_i = \sum_{i=1}^{I} p_{ij}$. The next step in the correspondence analysis leads to row and column profiles:

$$\mathbf{D_r^{-1}P} = \left[\frac{n_{hj}}{n_{h.}}\right] = \left[\frac{p_{hj}}{p_{h.}}\right] \tag{1}$$

$$\mathbf{D_c^{-1}P}^{\mathrm{T}} = \left[\frac{n_{hj}}{n_{.j}}\right] = \left[\frac{p_{hj}}{p_{.j}}\right] \tag{2}$$

where: $\mathbf{D_r} = diag(\mathbf{r})$ , and $\mathbf{D_c} = diag(\mathbf{c})$.

The problem of a graphical presentation co-occurrence of the categories of variables becomes more serious when the categorical variables are characterized by a large number of categories. For this purpose, the decomposition of **A** matrix by SVD should be applied, thanks to which it is possible to determine the coordinates of the categories of variables of interest, and to determine the degree of dispersion.

## 2. Singular value decomposition (SVD)

The first authors who focused on SVD were E. Beltrami (1873) and C. Jordan (1874). A further development of this method was proposed by A. Marshal and I. Olkin (1979). One of the most important works on SVD algorithm was presented by Eckart and Young in the first issue of *Psychometrika* (Eckart, Young, 1936). Psychometricians used this algorithm under name Eckart-Young decomposition. Other names include the basic structure (Horst, 1936, Green, Carroll, 1976), as well as the canonical form (Eckart, Young, 1936), or singular decomposition (Good, 1969, Kshirsagar, 1972). Today, it is known under the name of singular value decomposition. More information about this decomposition is given by J.M. Chambers (1977), K.R. Gabriel (1978), C.R. Rao (1980), and M. Greenacre and L.G. Underhill (1982). SVD in a correspondence analysis allows to determine the coordinates of points, which in turn allows for the application of the points representing the categories on the map perception.

The problem of decomposition of **A** matrix by SVD in a correspondence analysis have been addressed by R.A. Fisher (1940), M.J. Greenacre (1984), E.B. Anderson (1991), and J.D. Jobson (1992). The following discussion explains how decomposition of **A** matrix is connected to the value of inertia. An important element of this study is a graphical presentation of the configuration of points in a two-dimensional space for all the algorithms with the use of procedures in R.

### 2.1. SVD by Fisher

The method of distribution of **A** matrix by singular value was proposed by R.A. Fisher in 1940 (Borg, Groenen, 1997, van den Heijden, 1987). It is based on decomposition of **A** matrix according to the formula:

$$\mathbf{A} = \mathbf{U\Gamma V^T} \tag{3}$$

$$\mathbf{A} = \mathbf{D_r}^{-\frac{1}{2}}(\mathbf{N} - \mathbf{E})\mathbf{D_c}^{-\frac{1}{2}} \tag{4}$$

where: $\mathbf{E} = \mathbf{D_r 1 1^T D_c n^{-1}}$, or $e_{hj} = \frac{n_{h.} \times n_{.j}}{n}$, $\mathbf{D_r}$ is a diagonal matrix of values $n_h$, $\mathbf{D_c}$ is a diagonal matrix of values $n_{.j}$, **U** is row (left) eigenvectors ($H$ × $K$), **V** is column (right) eigenvectors ($J$ × $K$), **Γ** is a diagonal matrix of (positive) singular values in a descending order.

The principal coordinates of rows and columns are defined as:

$$\mathbf{F} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{U}n^{\frac{1}{2}} \tag{5}$$

$$\mathbf{G} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{V}n^{\frac{1}{2}} \tag{6}$$

The relationship between the singular values and the chi-square statistics is the following:

$$\lambda = tr\mathbf{\Lambda} = \frac{\chi^2}{n} = \sum_{k=1}^{K}\gamma_k^2 = \sum_{k=1}^{K}\lambda_k \tag{7}$$

where: $\lambda_k$ are eigenvalues of $\mathbf{A^T A}$ and $\mathbf{AA^T}$, $\gamma_k^2 = \lambda_k$ are squares singular value of **A**.

## 2.2. SVD by Greenacre

A decomposition method proposed by the differences in standardized residuals was proposed by M.J. Greenacre (1984). It presents **A** matrix as the product of three matrices:

$$\mathbf{A} = \mathbf{U\Gamma V}^{\mathrm{T}} \tag{8}$$

where: **Γ** is a diagonal matrix of singular values of **A** matrix in a descending order ($\gamma_1 > \gamma_2 > ... > \gamma_k$), **U** is row (left) eigenvectors, and **V** is column (right) eigenvectors.

Matrix **A** is defined as:

$$\mathbf{A} = \mathbf{D_r}^{-\frac{1}{2}}(\mathbf{P} - \mathbf{rc}^{\mathbf{T}})\mathbf{D_c}^{-\frac{1}{2}} \tag{9}$$

or

$$a_{hj} = \frac{p_{hj} - p_{h.}p_{.j}}{\sqrt{p_{h.}p_{.j}}} \tag{10}$$

Matrices **U** and **V** are orthogonal: $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. The following relation exists:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{U\Lambda U}^{\mathrm{T}} \tag{11}$$

$$\mathbf{AA}^{\mathrm{T}} = \mathbf{V\Lambda V}^{\mathrm{T}} \tag{12}$$

The principal coordinates of the rows and columns are defined as:

$$\mathbf{F} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{U\Gamma} \tag{13}$$

$$\mathbf{G} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{V\Gamma} \tag{14}$$

The relationship between the singular values and the chi-square statistics is the following:

$$\lambda = tr\mathbf{\Lambda} = tr\mathbf{A}^{\mathbf{T}}\mathbf{A} = tr\mathbf{AA}^{\mathbf{T}} = \sum_{k=1}^{K}\gamma_k^2 = \frac{\chi^2}{n} \tag{15}$$

## 2.3. SVD by Anderson

E.B. Anderson (1991) proposed an analysis starting with examining the difference between profiles and an average profile, setting a vector of the difference of profiles (for rows and columns, respectively) defined as:

$$\mathbf{h} = \left[ \frac{p_{h1}}{p_{h.}} - p_{.1}, \ldots, \frac{p_{hj}}{p_{h.}} - p_{.j} \right] \tag{16}$$

$$\mathbf{c} = \left[ \frac{p_{1j}}{p_{.j}} - p_{1.}, \ldots, \frac{p_{hj}}{p_{.j}} - p_{h.} \right] \tag{17}$$

If the differences in **c** and **h** vectors are equal to zero, then the variables are independent. E.B. Anderson (1991) proposed to define the decomposition as:

$$\frac{p_{hj}}{p_{h.}p_{.j}} - 1 = \sum_{k=1}^{K} \gamma_k u_{hk} v_{jk} \tag{18}$$

$$\mathbf{A} = \mathbf{U\Gamma V^T} \tag{19}$$

$$\mathbf{A} = \mathbf{D_r^{-1} P D_c^{-1}} - \mathbf{1_r 1_c^{-1}} = \mathbf{D_r^{-1} R D_c^{-1}} \tag{20}$$

where: $\mathbf{P} = \left[ p_{hj} \right]$, $\mathbf{R} = \left[ p_{hj} - p_{h.}p_{.j} \right]$, **U**, and **V** satisfy the condition: $\mathbf{U^T D_r U} = \mathbf{I} = \mathbf{V^T D_c V}$.

The principal coordinates of rows and columns are defined as:

$$\mathbf{F} = \mathbf{U\Gamma} \tag{21}$$

$$\mathbf{G} = \mathbf{V\Gamma} \tag{22}$$

The relationship between the singular values and the chi-square statistics is the following:

$$\lambda = tr\mathbf{\Lambda} = \frac{\chi^2}{n} = \sum_{k=1}^{K} \gamma_k^2 \tag{23}$$

## 2.4. SVD by Jobson

J.D. Jobson (1992) proposed the following decomposition of **A** matrix under SVD algorithm:

$$\mathbf{A} = \mathbf{U\Gamma V^T} \tag{24}$$

$$\mathbf{A} = (\mathbf{P} - \mathbf{rc^T}) \tag{25}$$

where: **r** – marginal frequencies of rows, and **c** – marginal frequencies of columns. Matrices **U** and **V** satisfy the condition:

$$\mathbf{U^T D_r^{-1} U} = \mathbf{I} = \mathbf{V^T D_c^{-1} V} \tag{26}$$

The principal coordinates of rows and columns are defined as:

$$\mathbf{F}=\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{U}\boldsymbol{\Gamma}=\mathbf{D}_{\mathbf{r}}^{-1}\left(\mathbf{P}-\mathbf{rc}^{\mathbf{T}}\right)\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{V} \tag{27}$$

$$\mathbf{G}=\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{V}\boldsymbol{\Gamma}=\mathbf{D}_{\mathbf{c}}^{-1}\left(\mathbf{P}-\mathbf{rc}^{\mathbf{T}}\right)\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{U} \tag{28}$$

The relationship between the singular values and the chi-square statistics is the following:

$$\lambda=tr\boldsymbol{\Lambda}=\frac{\chi^2}{n}=\sum_{k=1}^{K}\gamma_k^2 \tag{29}$$

In the next part of the paper, different approaches using different datasets are applied and compared.

## 3. Application and comparative analysis in R

The aim of this paper is to compare and visualize the available approaches to decomposition of **A** matrix with the use of SVD. In order to obtain a graphical presentation of the results in a perception map, the study is applied for real-life data based on a two-dimensional space.

To determine the coordinates of the categories in a two-dimensional space, and for a graphical presentation of the correspondence analysis under four different approaches to SVD, we applied our own procedures in *R*. In the study, we apply SVD function in a MASS package to visualize the results in a two-dimensional space. The graphs are intentionally left without a category's name to make them clear and transparent. It is also possible to use the individual author's procedures in *R* to make graphs in another option, using labels or colors.

In this study, we conduct a correspondence analysis based on the data on poverty and social exclusion from Eurostat (http://ec.europa.eu/eurostat). In a dataset, we included countries from the European Union in the period of 2008–2013: Belgium, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Slovenia, Slovakia, Finland, Sweden, UK, Iceland, and Norway.

## 4. Discussion

The correspondence analysis conducted in this study is based on fairly straightforward, classical results in the matrix theory. The central result is SVD, which is the basis of many multivariate methods, such as a principal component analysis, canonical correlation analysis, all forms of linear biplots, a discriminant analysis and metric multidimensional scaling. A graphical presentation of the results and a plot comparison is presented in Figure 1. The analysis of perception maps shows that, out of the four compared algorithms, there are two providing the same results, namely Fisher and Greenacre.

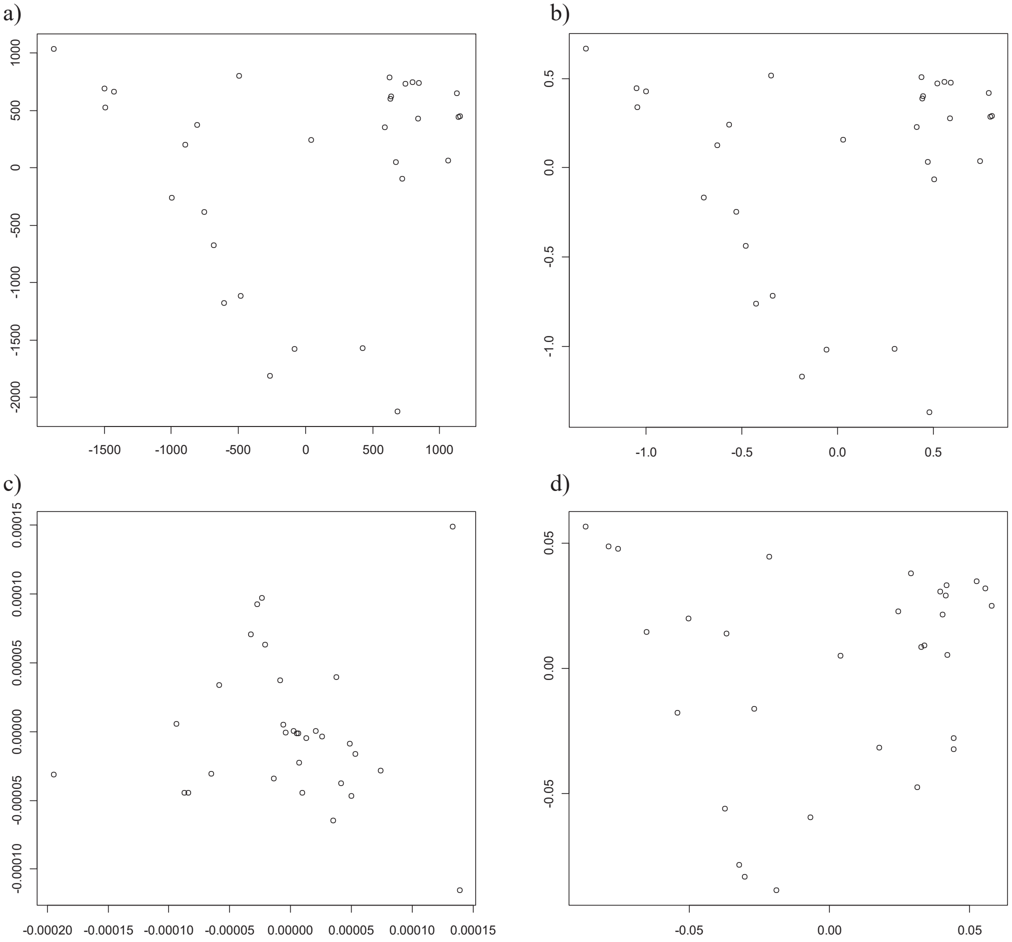

Figure 1. Graphical presentation of the correspondence analysis with the use of SVD algorithm by: (a) Fisher, (b) Greenacre, (c) Andersen, and (d) Jobson on the datasets from Eurostat.

Source: author's own calculations in *R*.

We also use another dataset on different financial founding sources in scientific disciplines (Greenacre 1993). The graphical results on a two-dimensional space are presented in Figure 2.

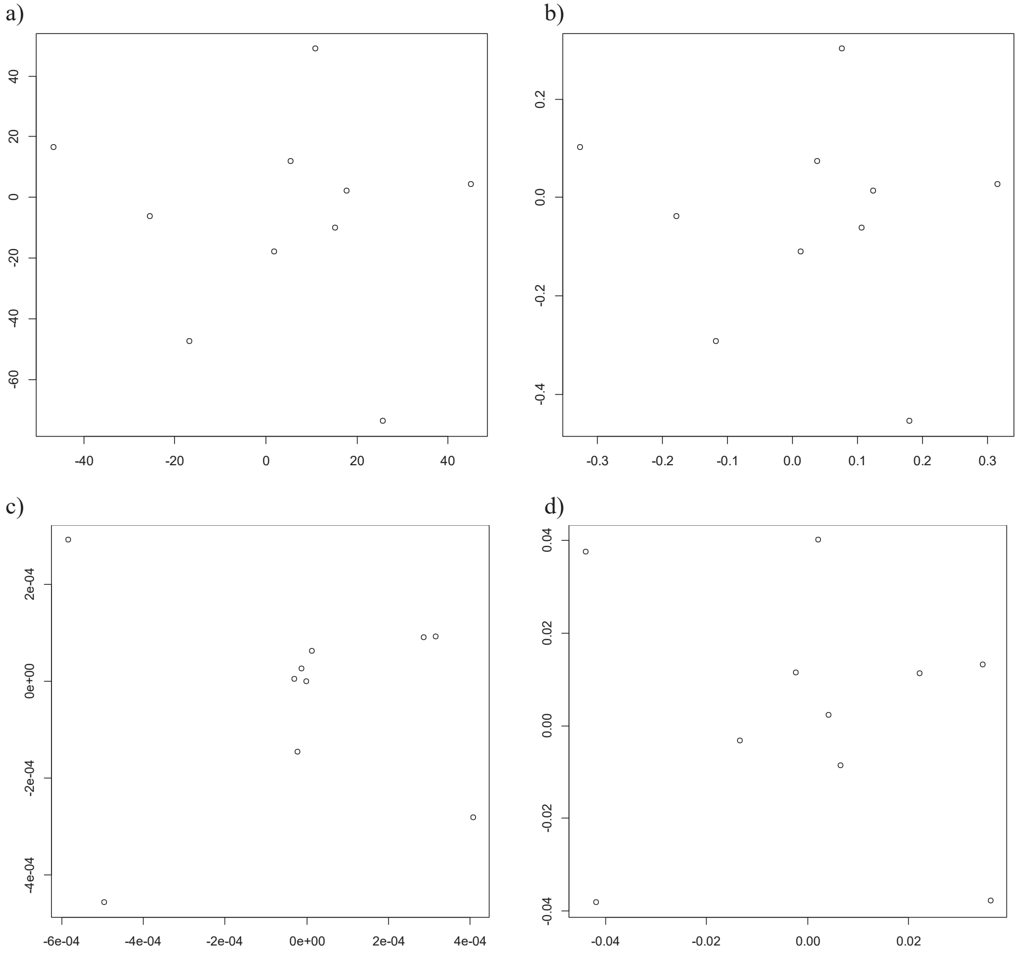

Figure 2. Graphical presentation of the correspondence analysis with the use of SVD algorithm by: (a) Fisher, (b) Greenacre, (c) Andersen, and (d) Jobson on the dataset from Greenacre (1993).

Source: author's own calculations in *R*.

The analysis of both datasets shows that there are two approaches which give the same graphical structure of points in a low-dimensional space, namely R.A. Fisher and M.J. Greenacre, however, the only difference is points coordinates.

SVD is the fundamental mathematical result for a correspondence analysis, as it is for other dimension reduction techniques, such as a principal component analysis, canonical correlation analysis, and a linear discriminant analysis. This matrix decomposition expresses any rectangular matrix as a product of three matrices of a simple structure.

Several similar studies on other data have been also conducted to prove that the same point configuration will be obtained for these approaches, and out of all the results for Fisher and Greenacre SVD algorithm, we obtain the same two-dimensional plot. However, Andersen's and Jobson's SVD approaches vary, resulting in different set points in a two-dimensional space.

## Conclusions

The paper presents four different algorithms of the distribution of matrix **A** with the use of SVD algorithm. They were presented by: R.A. Fisher (1940), M.J. Greenacre (1984), E.B. Anderson (1991), and J.D. Jobson (1992). In order to determine the points representing the categories of the variables in a two-dimensional space, we apply procedures in *R* software which allow to present graphically the results in a perception map.

A graphical presentation of the results for each algorithm proves that Fisher's and Greenacre's SVD algorithms are the same, however, Andersen's and Jobson's vary, resulting in a different set points in a two-dimensional space. The study has been also conducted for other datasets, giving the same result and the same point configuration for Fisher's and Greenacre's approaches.

The most common and popular method for SVD is Greenacre's, which is programmed in *R* automatically (SVD function in MASS package). The literature in the field of a correspondence analysis on the distribution of **A** matrix is quite poor, and there is a lack of systematization of the four presented algorithms. This work focused on the systematization of knowledge, comparison of SVD algorithms, and on presenting the results in a graphical form with the use of a correspondence analysis. The conducted study could encourage R-users and package authors to make a contribution to Fisher's work, and to make it available in *R*, since the results obtained with the use of Fisher's algorithm are the same as for Greenacre's. It would be also advisable to create SVD package in *R* that would be based on other approaches, just to make it possible to use computerized procedures for other approaches than Greenacre's, which is the most popular one and formalized in *R* packages for a correspondence analysis.

## Refferences

Anderson, E.B. (1991). *The statistical analysis of categorical data*. Berlin: Spinger-Verlag.

Beltrami, E. (1873). Sulle Funzioni Bilineari. *Giornale di Matematiche ud uso Degli Studenti Delle Universita*, *11*, 98–106.

Borg, I., Groenen, P. (1997). *Modern multidimensional scaling. Theory and application*. New York: Spinger-Verlag.

Chambers, J.M. (1977). *Computational methods for data analysis*. New York: Wiley.

Clausen, S.E. (1998). *Applied correspondence analysis. An introduction*. Thousand Oaks: Sage Publications.

Eckart, C., Young, G. (1936). The approximation of one matrix by an-other of lower rank. *Psychometrika*, *1*, 211–218.

Fisher R.A. (1940). The precision of discriminant functions. *Annals of Eugenics*, *10*, 422–429.

Gabriel, K.R. (1978). Least-squares approximation of matrices by additive and multiplicative models. *J. R. Statist. Soc. B*, *40*, 186–196.

Good, I.J. (1969). Some applications of the singular decomposition of a matrix. *Technometrics*, *11*, 823–831.

Green, P.E., Carroll, J.D. (1976). *Mathematical tools for applied multivariate analysis*. New York: Academic Press.

Greenacre, M., Underhill, L.G. (1982). Scaling a data matrix in low-dimensional Euclidean space. In: D.M. Hawkins, *Topics in applied multivariate analysis* (pp. 183–268). UK: Cambridge University Press.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M.J. (2010). *Biplots in Practice. Fundacion BBVA*.

Heijden VanDer, P.G.M. (1987). *Correspondence analysis of longitudinal categorical data*. Leiden: DSWO Press.

Horst, P. (1936). Obtaining a composite measure from a number of dif-ferent measures of the same attribute. *Psychometrika*, *1*, 53–60.

Jobson, J.D. (1992). *Applied multivariate data analysis Vol. II: Categorical and multivariate methods.* New York: Spinger-Verlag.

Jordan, C. (1874). Memoire sur les formes bilineaires. *Journal de Mathematiques Pures et Appliquees, Deuxieme Serie*, *19*, 37–39.

Korobeynikov, A, Larsen, R.M. (2016). *svd: Interfaces to Various State-of-Art SVD and Eigensolvers R package version 0.4.* Retrieved from: https://CRAN.R-project.org/package=svd.

Kshirsagar, A.M. (1972). *Multivariate analysis*. New York: Marcel Dekker.

Marshal, A., Olkin, I. (1979). *Inequalities: theory of majorization and its applications*. New York: Academic Press.

Rao, C.R. (1980). Matrix approximation and reduction of dimensional-ity in multivariate statistical analysis. In: P.R. Krishnaiah (ed.), *Multivariate analysis V* (pp. 3–22). North Holland, Amsterdam.

Stewart, G.W. (1993). On the Early History of the Singular Value Decomposition. *SIAM Review*, *4* (35), 551–566.

Qiu, Y., Mei, J., Guennebaud, G., Niesen, J., (2016). *RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems.R package version 0.12-0*. Retrieved from: https://CRAN.R-project.org/package=RSpectra.

Voronin, S., Martinsson, P.G. (2015). *RSVDPACK: Subroutines for Computing Partial Singular Value Decompositions via Randomized Sampling on Single Core, Multi Core, and GPU Architectures*. arXiv preprint (pp. 1–15). Retrieved from: http://arxiv.org/abs/1502.05366.