PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Justyna Brzezińska
Uniwersytet Ekonomiczny w Katowicach

ROZKŁAD MACIERZY WEDŁUG WARTOŚCI OSOBLIWYCH (SVD) W ANALIZIE KORESPONDENCJI

Streszczenie: Wizualizacja wyników analizy korespondencji jest możliwa dzięki dekompozycji macierzy według wartości osobliwych (*Singular Value Decomposition*). Celem niniejszej pracy jest przedstawienie różnych podejść i algorytmów metody SVD, stosowanych w analizie korespondencji, dzięki którym możliwe jest zaprezentowanie kategorii badanych zmiennych w jednym układzie odniesienia. W literaturze wymieniane są algorytmy metody SVD według czterech podejść: Fishera, Greenacre'a, Andersena, oraz Jobsona. Zaprezentowane zostaną także sposoby wyznaczania współrzędnych kategorii wierszowych oraz kolumnowych. Interesujący problem stanowi porównanie wszystkich podejść, gdyż w literaturze wykorzystywane jest głównie podejście zaproponowane przez Greenacre'a. W pracy zaprezentowano graficzne przedstawienie wyników każdej omawianej metody w programie **R**.

Słowa kluczowe: analiza korespondencji, rozkład według wartości osobliwych, SVD.

1. Wstęp

Analiza korespondencji jest metodą statystyki wielowymiarowej, zaliczaną do grupy metod badania współwystępowania zmiennych nominalnych. Graficzna prezentacja kategorii zmiennych w jednym układzie odniesienia jest możliwa dzięki dekompozycji macierzy różnic standaryzowanych **A** według wartości osobliwych (*Singular Value Decomposition* − SVD). Rozkład macierzy **A** na trzy macierze pozwala wyznaczyć współrzędne kategorii, dzięki którym możliwa jest graficzna prezentacja wyników, a także wartość inercji, będącej miarą rozproszenia punktów. W literaturze znane są cztery algorytmy rozkładu macierzy **A** według wartości osobliwych: Fishera [1940], Greenacre'a [1982], Andersena [1991] oraz Jobsona [1992]. Celem niniejszej pracy jest usystematyzowanie i porównanie wyników uzyskanych dzięki zastosowaniu czterech zaprezentowanych algorytmów. Literatura polska, a także światowa, jest dość uboga, jeśli chodzi o usystematyzowanie wiedzy z zakresu metody SVD, wykorzystywanej w analizie korespondencji. Nie-

wiele można znaleźć porównań pomiędzy istniejącymi algorytmami oraz jej wynikami. W pierwszej części pracy zaprezentowano algorytm oraz podstawowe pojęcia analizy korespondencji. W drugiej zaś przedstawiono cztery sposoby dekompozycji macierzy **A** według wartości osobliwych, a także sposoby wyznaczania współrzędnych punktów reprezentujących kategorie zmiennych. W pracy wykorzystano autorskie procedury w programie **R**, dzięki którym zaprezentowano i porównano graficzną konfigurację punktów w dwuwymiarowej przestrzeni.

2. Algorytm analizy korespondencji

Pierwszym etapem analizy korespondencji jest budowa macierzy kontyngencji (tablica liczebności) N. Na jej podstawie wyznaczana jest macierz korespondencji **P**. Częstości brzegowe wierszy nazywane są masami wierszy, a częstości brzegowe kolumn − masami kolumn. Elementy $p_{h.}$ i $p_{.j}$ tworzą wektory częstości brzegowych wierszy r oraz kolumn c. Kolejnym krokiem w analizie korespondencji jest wyznaczenie profili wierszy i profili kolumn według wzorów:

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

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

Problem graficznej prezentacji jednoczesnego występowania kategorii obu zmiennych staje się poważniejszy, gdy zmienne charakteryzują się dużą liczbą kategorii. W tym celu zastosowanie znajduje rozkład macierzy **A** według wartości osobliwych, dzięki której możliwe jest wyznaczenie współrzędnych kategorii zmiennych oraz określenie stopnia ich rozproszenia.

3. Rozkład macierzy według wartości osobliwych SVD

Pierwszymi twórcami metody rozkładu macierzy według wartości osobliwych (SVD) byli Beltrami i Jordan. Dalszy rozwój metoda ta zawdzięcza Marshalowi i Olkinowi [1979]. Jedną z ważniejszych prac na temat metody SVD zaprezentowali Eckart i Young w pierwszym numerze „Psychometriki" [1936]. Psychometrycy używają nazwy metody *dekompozycja Eckart-Young* (*Eckart-Young decomposition*). Inne nazwy to *struktura bazowa* (*basic structure*) (Horst [1936], Green, Carroll [1976]), a także *forma kanoniczna* (*canonical form*) (Eckart, Young [1936]; Johnson [1963]) lub *dekompozycja osobliwa* (*singular decomposition*) (Good [1969], Kshirsagar [1972]). Dziś powszechna jest nazwa rozkład macierzy według

wartości osobliwych (SVD). Szerzej temat ten opisują Chambers [1977], Gabriel [1978], Rao [1980] oraz Greenacre i Underhill [1982]. Rozkład SVD w przypadku analizy korespondencji umożliwia wyznaczenie współrzędnych punktów, co w konsekwencji pozwala na naniesienie punktów reprezentujących kategorie na mapę percepcji.

Rozwiązania problemu dekompozycji macierzy **A** według wartości osobliwych w analizie korespondencji zostały opracowane przez Fishera [1940], Greenacre'a [1984], Andersena [1991] oraz Jobsona [1992]. W poniższych rozważaniach przedstawiono sposoby dekompozycji macierzy **A** oraz zdefiniowano pojęcie inercji. Istotnym elementem badania jest zaprezentowanie graficznej konfiguracji punktów w dwuwymiarowej przestrzeni dla wszystkich algorytmów.

3.1. Rozkład macierzy według wartości osobliwych Fishera

Metoda rozkładu macierzy według wartości osobliwych zaproponowana przez Fishera w 1940 roku [Borg, Groenen 1997; van der Heijden 1987] polega na dekompozycji macierzy **A** według wzoru:

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

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

gdzie: $\mathbf{E} = \mathbf{D_r 11^T D_c n^{-1}}$ lub $e_{hj} = \dfrac{n_{h.} \cdot n_{.j}}{n}$,

$\mathbf{D_r}$ – macierz diagonalna o wartościach $n_{h.}$,

$\mathbf{D_c}$ – macierz diagonalna o wartościach $n_{.j}$,

$\mathbf{U}$ – macierz lewych wektorów osobliwych $(H \times K)$,

$\mathbf{V}$ – macierz prawych wektorów osobliwych $(J \times K)$,

$\mathbf{\Gamma}$ – macierz diagonalna niezerowych wartości osobliwych γ_k $(k = 1, 2, ..., K)$ macierzy $\mathbf{A}$ $(K \times K)$ uporządkowanych malejąco.

Współrzędne dla kategorii w wierszach oraz kolumnach wyznaczane są według wzorów:

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

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

Pomiędzy wartościami osobliwymi γ_k a statystyką χ^2 zachodzi zależność:

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

gdzie: λ_k – wartości własne macierzy $\mathbf{A^T A}$ oraz $\mathbf{AA^T}$,
$\gamma_k^2 = \lambda_k$ – kwadraty wartości osobliwych macierzy **A**.

3.2. Rozkład macierzy według wartości osobliwych Greenacre'a

Metoda rozkładu macierzy różnic standaryzowanych zaproponowana przez Greenacre'a [1984] polega na zaprezentowaniu macierzy **A** w postaci iloczynu trzech macierzy [Stanimir 2005]:

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

gdzie: $\mathbf{\Gamma}$ – macierz diagonalna o wymiarach $(K \times K)$, utworzona z wartości osobliwych macierzy **A**, uporządkowanych w sposób malejący wartości osobliwych ($\gamma_1 > \gamma_2 > ... > \gamma_k$, $k = 1, 2, ..., K$),

U – macierz lewych wektorów osobliwych macierzy $\mathbf{A^T A}$, odpowiadających wartościom własnym $\gamma_1^2, \gamma_2^2, ..., \gamma_k^2$,

V – macierz prawych wektorów osobliwych macierzy $\mathbf{AA^T}$, odpowiadających wartościom własnym $\gamma_1^2, \gamma_2^2, ..., \gamma_k^2$.

Macierz **A** zdefiniowana jako:

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

jest macierzą różnic standaryzowanych o wymiarach $H \times J$, gdzie $a_{hj} = \frac{p_{hj} - p_{h.}p_{.j}}{\sqrt{p_{h.}p_{.j}}}$.

Macierze **U** oraz **V** są ortogonalne ($\mathbf{U^T U} = \mathbf{V^T V} = \mathbf{I}$). W celu znalezienia lewych i prawych wektorów osobliwych korzysta się z rozkładu macierzy według wartości własnych, gdyż kolumny macierzy **U** są wektorami własnymi macierzy $\mathbf{A^T A}$, natomiast macierz **V** zawiera w kolumnach wektory własne macierzy $\mathbf{AA^T}$. Wartości osobliwe macierzy **A** są pierwiastkami kwadratowymi wartości własnych macierzy $\mathbf{A^T A}$ oraz $\mathbf{AA^T}$. Zachodzą zatem zależności:

$$\mathbf{A^T A} = \mathbf{U\Lambda U^T}, \tag{10}$$

$$\mathbf{A}\mathbf{A}^{\mathbf{T}} = \mathbf{V}\Lambda\mathbf{V}^{\mathbf{T}}, \tag{11}$$

gdzie $\Lambda = \mathbf{\Gamma}^2$, a $\Gamma = [\lambda_k]$, λ_k to wartości własne macierzy $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ oraz $\mathbf{A}\mathbf{A}^{\mathbf{T}}$. Wartości własne macierzy $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ to takie, dla których zachodzi równanie:

$$\det(\mathbf{A}^{\mathbf{T}}\mathbf{A} - \lambda_k\mathbf{I}) = 0. \tag{12}$$

Pomiędzy wartościami osobliwymi a wartością statystyki chi-kwadrat zachodzi zależność:

$$\lambda = tr\Lambda = tr\mathbf{A}^{\mathbf{T}}\mathbf{A} = tr\mathbf{A}\mathbf{A}^{\mathbf{T}} = \sum_{k=1}^{K}\gamma_k^2 = \frac{\chi^2}{n}. \tag{13}$$

Współrzędne dla kategorii wierszowej wyznaczane są według formuły:

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{U}\mathbf{\Gamma}, \tag{14}$$

gdzie $\mathbf{D}_{\mathbf{r}}$ to diagonalna macierz częstości brzegowych wierszy macierzy **P**.

Współrzędne dla kategorii zapisanych w kolumnach tablicy kontyngencji wyznaczane są według formuły:

$$\mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{V}\mathbf{\Gamma}, \tag{15}$$

gdzie $\mathbf{D}_{\mathbf{c}}$ to diagonalna macierz częstości brzegowych kolumn macierzy **P**.

3.3. Rozkład macierzy według wartości osobliwych Andersena

Andersen [1991] rozpoczyna analizę od badania różnicy profili od profilu średniego, wyznaczając wektor różnic profili (odpowiednio dla wierszy i kolumn) jako:

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

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

Jeśli wartości w wektorach różnic **h** i **c** są równe zeru, to cechy są niezależne. Andersen proponuje, by dokonać dekompozycji wyrażenia postaci:

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

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

$$\mathbf{A} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\mathbf{D}_{\mathbf{c}}^{-1} - \mathbf{1}_{\mathbf{r}}\mathbf{1}_{\mathbf{c}}^{-1} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{R}\mathbf{D}_{\mathbf{c}}^{-1}, \tag{20}$$

gdzie $\mathbf{P} = \left[p_{hj}\right]$, $\mathbf{R} = \left[p_{hj} - p_{h.}p_{.j}\right]$ oraz **U** i **V** spełniają warunek: $\mathbf{U}^{\mathbf{T}}\mathbf{D}_{\mathbf{r}}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathbf{T}}\mathbf{D}_{\mathbf{c}}\mathbf{V}$. Współrzędne kategorii odpowiednio w wierszach oraz kolumnach wyznacza się ze wzorów:

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

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

Między wartościami osobliwymi macierzy **A** a statystyką χ^2 zachodzi zależność:

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

3.4. Rozkład macierzy według wartości osobliwych Jobsona

Jobson [1992] zaproponował sposób dekompozycji macierzy **A** według wartości osobliwych zgodnie z formułą:

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

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

gdzie: **r** – częstości brzegowe wierszy,
c – częstości brzegowe kolumn,
P – macierz częstości zaobserwowanych o elementach $\left[p_{hj}\right]$.

Macierze **U** oraz **V** spełniają zależność: $\mathbf{U}^{\mathbf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{U} = \mathbf{I} = \mathbf{V}^{\mathbf{T}}\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{V}$. Współrzędne kategorii występujących w wierszach wyznaczane są według formuły:

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{U\Gamma} = \mathbf{D}_{\mathbf{r}}^{-1}\left(\mathbf{P} - \mathbf{rc}^{\mathbf{T}}\right)\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{V}, \tag{26}$$

natomiast dla kategorii kolumnowych według formuły:

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

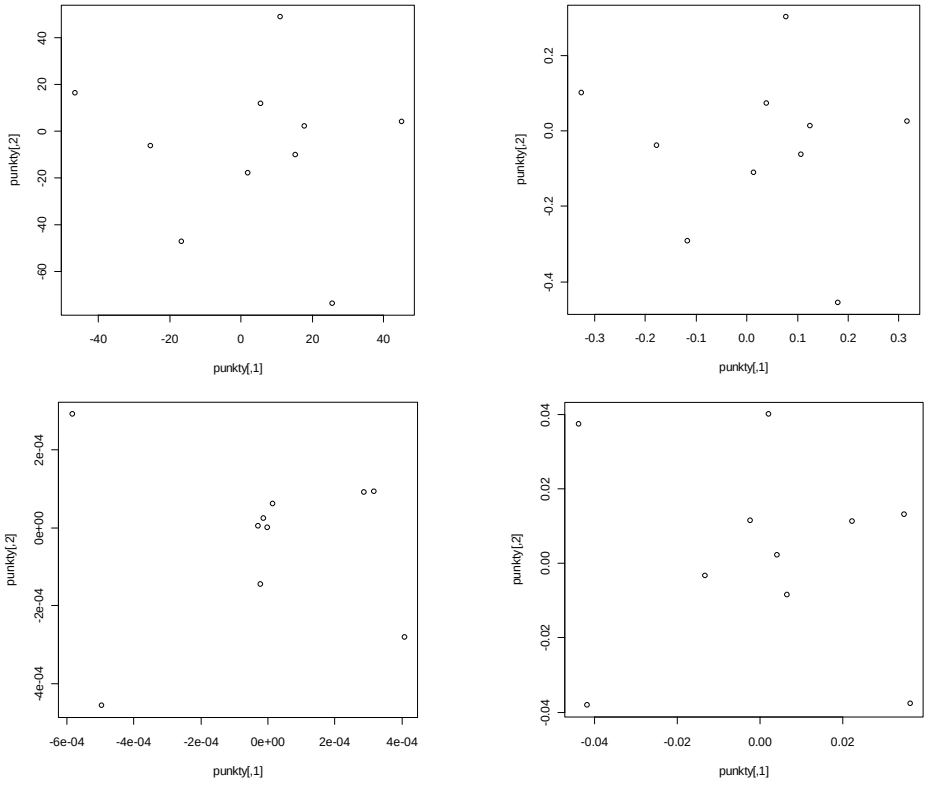
Pomiędzy statystyką χ^2, a wartościami osobliwymi zachodzi następująca zależność:

$$\lambda = tr\mathbf{\Lambda} = \frac{\chi^2}{n} = \sum_{k=1}^{K} \gamma_k^2 , \tag{28}$$

gdzie $\mathbf{\Lambda} = \mathbf{\Gamma}^2$ ($\mathbf{\Gamma} = [\lambda_k]$, λ_k – wartości własne odpowiednio macierzy $\mathbf{A^T A}$ oraz $\mathbf{A A^T}$).

4. Graficzna prezentacja wyników w przestrzeni dwuwymiarowej

Celem zaprezentowanych podejść rozkładu macierzy **A** według wartości osobliwych jest graficzna prezentacja wyników w postaci mapy percepcji. W niniejszym badaniu analizę korespondencji przeprowadzono na danych dotyczących źródeł finansowania badań naukowych w różnych dyscyplinach [Greenacre 1993]. Do wyznaczenia współrzędnych kategorii w dwuwymiarowej przestrzeni wykorzystano



Rys. 1. Graficzna prezentacja punktów otrzymana w wyniku zastosowania rozkładu macierzy według Fishera, Greenacre'a, Andersena oraz Jonsona

Źródło: opracowanie własne w programie **R**.

autorskie procedury w programie **R**. Wyznaczone współrzędne pozwoliły na zaprezentowanie konfiguracji punktów w przestrzeni niskowymiarowej, dzięki czemu możliwe jest porównanie wyników czterech algorytmów. Graficzną prezentacją wyników dla każdej omówionej metody jest mapa percepcji w przestrzeni dwuwymiarowej (rys. 1).

Wniosek z przeprowadzonej analizy jest taki, że dwa pierwsze algorytmy, tj. metoda Fishera oraz Greenacre'a są zbieżne, dają identyczną konfigurację punktów. Algorytmy Jobsona i Andersena dają odmienne wyniki, co widoczne jest na mapie percepcji.

5. Podsumowanie

W artykule zaprezentowane zostały cztery algorytmy rozkładu macierzy różnic standaryzowanych **A** według wartości osobliwych (SVD): Fishera [1940], Greenacre'a [1984], Andersena [1991] oraz Jobsona [1992]. Do wyznaczenia punktów reprezentujących kategorie zmiennych w przestrzeni dwuwymiarowe zaprezentowane zostały autorskie procedury w programie **R**. Graficzna prezentacja wyników dla każego algorytmu dowodzi, że dwa pierwsze spośród czterech zaprezentowanych podejść, tj. algorytm Fishera oraz Greenacre'a są zbieżne, czego rezultatem jest identyczna konfiguracja punktów na mapie percepcji. Dwa kolejne algorytmy, tj. Andersena oraz Jobsona, różnią się, czego rezultatem jest odmienna konfiguracja punktów. Najczęściej wykorzystywane spośród czterech zaprezentowanych algorytmów jest podejście Greenacre'a, oprogramowane w programie **R** (funkcja `svd()` w pakiecie `MASS`). Literatura z zakresu analizy korespondencji na temat sposobów rozkładu macierzy **A** jest dosyć uboga i brak w niej systematyzacji dotyczącej czterech zaprezentowanych algorytmów. W literaturze polskiej dostępna jest monografia autorstwa Stanimir [2005]. Niniejsza praca stanowi usystematyzowanie tej wiedzy oraz zaprezentowanie wyników w postaci graficznej.

Literatura

Andersen E.B., *The Statistical Analysis of Categorical Data*, Spinger-Verlag, Berlin 1991.

Borg I., Groenen P., *Modern Multidimensional Scaling. Theory and Application*, Spinger-Verlag, New York 1997

Chambers J.M., *Computational Methods for Data Analysis*, Wiley, New York 1977.

Clausen S.E., *Applied Correspondence Analysis. An Introduction*, Sage Publications, Thousand Oaks 1998.

Eckart C., Young G., *The approximation of one matrix by another of lower rank*. „Psychometrika" 1936, 1, s. 211-218.

Fisher R.A., *The precision of discriminant function*, „Annals of Eugenics" 1940, 10, s. 422-429.

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

Good I.J., *Some applications of the singular decomposition of a matrix*, „Technometrics" 1969, 11, s. 823-831.
Green P.E., Carroll J.D., *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York 1976.
Greenacre M.J., *Correspondence Analysis in Practice*, Academic Press, London 1993.
Greenacre M.J., *Theory and Applications of Correspondence Analysis, Academic Press*, London 1984.
Greenacre M.J., Underhill L.G., *Scaling a Data Matrix in Low-Dimensional Euclidean Space*, [w:] D.M. Hawkins (red.), *Topics in Applied Multivariate Analysis*, Cambridge University Press, Cambridge 1982.
Heijden P.G.M. van der, *Correspondence Analysis of Longitudinal Categorical Data*, DSWO Press, Leiden 1987.
Horst P., *Obtaining a composite measure from a number of different measures of the same attribute*, „Psychometrika" 1936, 1, s. 53-60.
Jobson J. D., *Applied Multivariate Data Analysis*, vol. II: *Categorical and Multivariate Methods*, Spinger-Verlag, New York 1992.
Kshirsagar A.M., *Multivariate Analysis*, Marcel Dekker, New York 1972.
Marshal A., Olkin I., *Inequalities*: *Theory of Majorization and Its Applications*, Academic Press, New York 1979.
Rao C.R., *Matrix approximation and reduction of dimensionality in multivariate statistical analysis*, [w:] R.P. Krishnaiah (red.), Multivariate analysis, Amsterdam 1980.
Stanimir A., *Analiza korespondencji jako narzędzie badania zjawisk ekonomicznych*, Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław 2005.
Walesiak M., Gatnar E., *Statystyczna analiza danych z wykorzystaniem programu R*, Wyd. Naukowe PWN, Warszawa 2009.

SINGULAR VALUE DECOMPOSITION (SVD) IN CORRESPONDENCE ANALYSIS

Summary: Correspondence analysis is a multivariate descriptive method that allows to visualize data as points in low-dimensional space. Singular value decomposition (*SVD*) is a method used for matrix decomposition. There are four approaches in *SVD* proposed by Fisher [1940], Greeancre [1981], Andersen [1991] and Jobson [1992]. A graphical result is presented in two-dimensional space. The most popular and widely used approach is proposed by Greenacre. This paper compares four approaches with its graphical presentation with the use of **R** software.