

Andrzej Dudek

Akademia Ekonomiczna we Wrocławiu

ANALIZA CZYNNIKOWA OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

W wielu współczesnych zagadnieniach klasyfikacyjnych tradycyjne podejście do reprezentowania danych w postaci macierzy lub wielowymiarowej tablicy liczb przestaje być wystarczające, a coraz więcej autorów postuluje o uwzględnianie w badaniach również danych nie mających reprezentacji numerycznej (zob. np. [Analysis of... 2000]). Jednym z rodzajów wielowymiarowej analizy statystycznej wychodzącym naprzeciw temu postulatowi jest symboliczna analiza danych, będąca jej rozszerzeniem na dane reprezentowane w postaci tablic obiektów symbolicznych. Każda komórka w tablicy obiektów symbolicznych (obserwacja na zmiennej symbolicznej) może zawierać dane w postaci pojedynczych wartości liczbowych, przedziałów liczbowych, zbiorów danych, zbiorów danych z wagami lub wartości tekstowych (jakościowych). Symboliczna analiza danych wprowadza nowe metody przetwarzania takich danych lub implementuje znane metody, tak aby mogły operować na danych symbolicznych.

W artykule przedstawiono rozszerzenie metod analizy czynnikowej, szczególnie metody głównych składowych, na dane reprezentowane w postaci obiektów symbolicznych. Omówiono algorytmy *Vertices* i *Centers* przeprowadzające analizę czynnikową metodą głównych składowych na danych reprezentowanych w postaci obiektów symbolicznych. Przeprowadzono analizę głównych składowych dla danych dotyczących rynku polskich biur podróży w latach 2000-2002, traktując dane wejściowe jako tabelę danych symbolicznych.

2. Obiekt symboliczny i zmienna symboliczna

Według E. Gatnara [1998, s. 62] „termin taksonomia symboliczna określa grupę metod klasyfikacji bezwzorcowej charakteryzujących się tym, że ich przedmiotem

są przede wszystkim obiekty symboliczne, których cechy mają charakter jakościowy, tj. są reprezentowane przez zmienne mierzone na skalach słabych: nominalnej i porządkowej. Podobieństwo tych obiektów jest ujmowane przez miary o charakterze heurystycznym, wykorzystujące teorię informacji, metody statystyczne itp. Powstałe w efekcie klasyfikacji skupienia są łatwiejsze do interpretacji dzięki zastosowaniu do ich opisu koniunkcji cech należących do nich obiektów.”.

Istotną cechą odróżniającą obiekty symboliczne od danych używanych w numerycznych metodach statystycznych jest odejście od postulatu, aby poszczególne zmienne były reprezentowane przez pojedyncze wartości liczbowe. Cecha obiektu symbolicznego może przyjmować więcej niż jedną wartość, może reprezentować przedział liczbowy, rozkład z odpowiednimi wagami czy dane w postaci nienumerycznej.

O ile typowa postać danych dla większości metod statystycznej analizy wielowymiarowej jest zgodna z postacią danych w tab. 1, o tyle tabela obiektów symbolicznych może być (jak w tab. 2) znacznie bardziej złożona, zawierając dane w postaci typowej dla zmiennych symbolicznych.

Tabela 1. Dane w postaci „tradycyjnej”

	Zmienna 1	Zmienna 2	Zmienna 3
1	1	108	11,98
2	1,3	123	-23,37
3	0,9	99	14,35
...	...	...	...

Źródło: opracowanie własne.

Tabela 2. Tabela obiektów symbolicznych

	Zmienna 1	Zmienna 2	Zmienna 3	Zmienna 4
1	(0,9; 1,0)	{106, 108, 110}	11,98	{niebieski, zielony}
2	(1; 2)	{123, 124, 125}	23,37	{jasno-szary}
3	(0,9; 1,3)	{100, 102, 99, 97}	14,35	{pastelowy}
...	...	...	...	...

Źródło: opracowanie własne na podstawie [Diday 2002].

Obserwacje na zmiennych w obiektach symbolicznych mogą zawierać [Diday 2002]:

- dane numeryczne, np. *przebieg* = 12347,
- dane tekstowe (jakościowe), np. *marka* = Renault,
- przedziały liczbowe, np. *spalanie* = [7,10],
- dane w postaci listy wartości, np. *wyposażenie* = {ABS, klimatyzacja, czujniki parkowania},
- dane w postaci listy wartości z wagami, np. *kolor* = {22% czerwony, 43% czarny, 35% niebieski}.

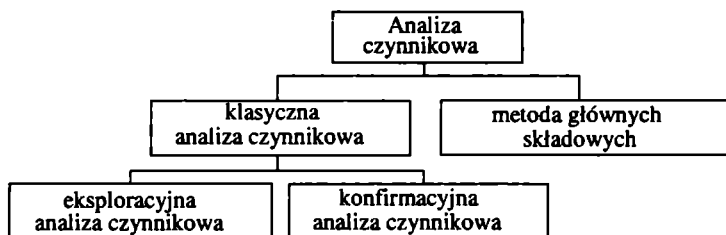
Ponadto zmienne w obiektach symbolicznych mogą być [Diday 2002]:

- taksonomiczne – reprezentujące strukturę hierarchiczną,
- hierarchicznie zależne,
- związane zależnościami logicznymi.

Dodatkowo wśród danych w obiektach symbolicznych może występować wartość *N/A* oznaczająca, że wartość tej zmiennej nie jest określona.

3. Analiza czynnikowa i metoda głównych składowych

Metoda głównych składowych, opracowana przez H. Hotellinga w 1933 r., pozwala na przedstawienie danych z przestrzeni $m > 3$ wymiarowej w przestrzeni 2- lub 3-wymiarowej. Jej miejsce wśród metod analizy czynnikowej przedstawia rys. 1.



Rys 1. Analiza czynnikowa i metoda głównych składowych
Źródło: na podstawie [Metody statystycznej... 2004, s. 186-232].

Matematyczny model analizy czynnikowej można zapisać jako (por. np. [Metody statystycznej... 2004, s. 186-232; Jajuga 1993, s. 210-214]):

$$\mathbf{Z} = \mathbf{B}\mathbf{S}, \quad (1)$$

gdzie: $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_m]^T$ – macierz standaryzowanych zmiennych,

$\mathbf{Z}_j = (z_{1j}, z_{2j}, \dots, z_{nj})$, $i \in \{1, \dots, n\}$ – numer obserwacji,

$\mathbf{B} = [b_{ip}]_{m \times m}$ – macierz współczynników (ładunków) głównych składowych,

$\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_m]^T$ – macierz głównych składowych, $\mathbf{S}_p = (s_{p1}, s_{p2}, \dots, s_{pn})$,

$j \in \{1, 2, \dots, m\}$ – numer zmiennej,

$p \in \{1, 2, \dots, m\}$ – numer głównej składowej,

Ładunki czynnikowe w metodzie głównych składowych określa się w sposób iteracyjny. Najpierw ustala się ładunki pierwszego czynnika F_1 poprzez maksymalizację udziału tego czynnika w wariancji wszystkich zmiennych (W_1), tzn. maksymalizując funkcję:

$$W_1 = \sum_{p=1}^m b_{p1}^2. \quad (2)$$

Następnie oblicza się macierz pozostałości korelacyjnej:

$$\mathbf{R}_1 = \mathbf{R} - \mathbf{B}_1 \mathbf{B}_1^T, \quad (3)$$

gdzie: $\mathbf{R}$ – macierz korelacji zmiennych,

$\mathbf{R}_1$ – zmodyfikowana macierz korelacji,

a w miejsce $\mathbf{B}_1 = [b_{p1}]$, $p \in \{1, 2, \dots, m\}$ podstawia się wartości ładunków pierwszego czynnika. Określona w ten sposób $\mathbf{R}_1$ służy do wyznaczenia ładunku drugiego czynnika głównego F_2 . W analogiczny sposób wyznacza się ładunki trzeciego czynnika i głównych czynników następnych do otrzymania wymaganego, łącznego stopnia wyjaśniania przez nie wariancji zmiennych lub dopóki udział kolejnego czynnika jest nie mniejszy niż wcześniej ustalona wartość lub dla z góry określonej liczby głównych składowych.

W celu obliczenia ekstremum warunkowego funkcji (2) wyznacza się wszystkie niezerowe i uporządkowane malejąco wartości własne macierzy korelacji $\mathbf{R}$ (λ_j , $j \in \{1, 2, \dots, m\}$) oraz odpowiadające im wektory własne ($\mathbf{V}_j = [v_{jp}]$, $p \in \{1, 2, \dots, m\}$).

Do obliczenia ładunku czynnikowego p -tego czynnika głównego j -tej zmiennej służy wzór:

$$b_{jp} = \frac{\sqrt{\lambda_p} v_{jp}}{\sqrt{\sum_{j=1}^m v_{jp}^2}}. \quad (4)$$

Procentowy zasób zmienności wspólnej wyjaśniony przez j -ty czynnik oblicza się ze wzoru:

$$\hat{h}_p = \frac{\lambda_p}{\sum_{j=1}^m \lambda_j} \cdot 100\% = \frac{\lambda_p}{m} \cdot 100\%. \quad (5)$$

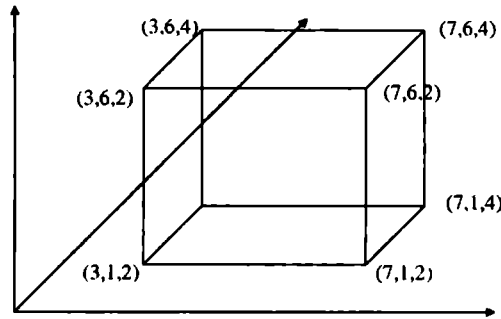
To, iż obliczenie funkcji określonej we wzorze (2) odbywa się poprzez obliczenie wartości i wektorów własnych macierzy danych, implikuje opisanie w uwagach końcowych, konsekwencje dla stosowania metody głównych składowych dla danych symbolicznych.

4. Algorytmy metody głównych składowych dla obiektów symbolicznych

Algorytmy *Vertices* i *Centers* pozwalają na zastosowanie metody głównych składowych dla obiektów symbolicznych, zawierających zmienne w postaci przedziałów liczbowych. Pierwszy z nich opiera się na zamianie reprezentacji danych w postaci przedziałów na wierzchołki (*Vertices*) hiperprostokątów, których krawędzie odpowiadają tym przedziałom, drugi – na zamianie reprezentacji tych danych na środki ciężkości (*Centers*) hiperprostokątów.

4.1. Algorytm *Vertices*

Założmy, że obiekt symboliczny składa się z trzech zmiennych reprezentujących przedziały liczbowe (3, 7); (1, 6); (2, 4). Taki obiekt symboliczny można traktować jako hiperprostopadłościan, zgodnie z rys. 2.



Rys 2. Reprezentacja danych w przestrzeni 3-wymiarowej dla algorytmu *Vertices*

Źródło: opracowanie własne na podstawie [Analysis of ... 2000].

Nie tracąc ogólności zapisu, możemy jednoznacznie przedstawić hiperprostopadłościan w postaci macierzy M_H :

$$M_H = \begin{bmatrix} 3 & 1 & 2 \\ 3 & 1 & 4 \\ 3 & 6 & 2 \\ 3 & 6 & 4 \\ 7 & 1 & 2 \\ 7 & 1 & 4 \\ 7 & 6 & 2 \\ 7 & 6 & 4 \end{bmatrix}. \quad (6)$$

Kolejne wiersze odpowiadają wierzchołkom hiperprostopadłościanu.

Taki sposób zamiany danych jest wykorzystany w algorytmie *Vertices*, składającym się z czterech bloków (według [Analysis of ... 2000, s. 202-204]):

1. Każdy obiekt symboliczny, składający się z p zmiennych reprezentujących przedziały, jest zamieniany na macierz o $n \cdot 2^m$ wierszach i m kolumnach, zawierającą wierzchołki hiperprostopadłościanu wyznaczonego przez przedziały.

2. Zamiast tabeli n obiektów symbolicznych jest tworzona macierz o wymiarach $n \cdot 2^m \times n$, łącząca macierze otrzymane w punkcie 1.

3. Dla danych z punktu 2 stosuje się metodę głównych składowych, wybierając wymiar $s < p$.

4. Z otrzymanych w punkcie 3 głównych składowych zostają skonstruowane główne składowe reprezentowane w postaci przedziałów. Wartość składowej

głównej reprezentowanej w postaci przedziału dla v -tej zmiennej i -tego obiektu wyraża się wzorem:

$$y_{iv} = [\underline{y}_{iv}, \overline{y}_{iv}], \quad (7)$$

gdzie: $\underline{y}_{iv} = \min_{k \in L_i}(y_{kv})$,

$\overline{y}_{iv} = \max_{k \in L_i}(y_{kv})$,

L_i – zbiór wszystkich indeksów macierzy z punktu 2 odpowiadających i -temu obiektowi symbolicznemu.

4.2. Algorytm *Centers*

Rozmiar macierzy wykorzystywanej przez algorytm *Vertices* rośnie wykładniczo w zależności od liczby zmiennych zawartych w obiekcie symbolicznym (m). Złożoność pamięciowa (rozmiar danych w pamięci komputera niezbędny do realizacji algorytmu), a co za tym idzie, złożoność obliczeniowa (gdyż metoda składowych głównych wymaga dla minimalizacji funkcji (1) obliczenia wartości i wektorów własnych tej macierzy a czas tej czynności jest wprost proporcjonalny do rozmiaru macierzy) algorytmu jest wykładnicza. Stąd dla obiektów symbolicznych, zawierających znaczną liczbę zmiennych, lepszym rozwiązaniem jest stosowanie uproszczonej jego wersji, czyli algorytmu *Centers*. Schemat działania algorytmu *Centers* można zapisać w trzech punktach (według [Analysis of ... 2000, s. 205-206]):

1. Przekształcamy tablicę danych symbolicznych na macierz zawierającą środki ciężkości hiperprostopadłościanów.

2. Dla danych z punktu 1 stosujemy metodę składowych głównych wybierając wymiar $s < m$.

3. Z otrzymanych w punkcie 2 głównych składowych konstruujemy główne składowe reprezentowane w postaci przedziałów.

Algorytmy *Vertices* i *Centers* pozwalają na zastosowanie metody głównych składowych dla obiektów symbolicznych. Przykład takiego wykorzystania tych metod znajduje się w następnej części artykułu.

5. Zastosowania metody głównych składowych dla obiektów symbolicznych

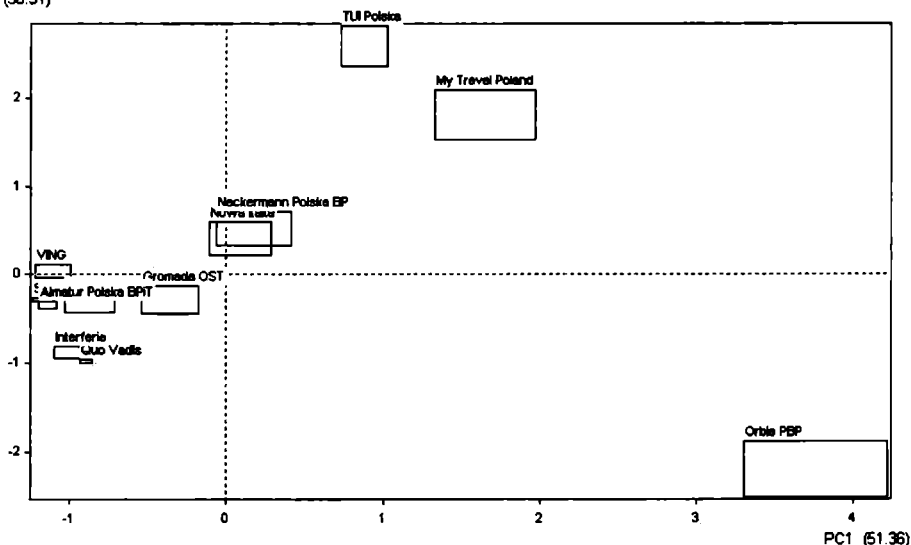
Tabela 3 zawiera dane o rynku biur podróży w Polsce w latach 2000-2002, przedstawione w postaci przedziałów liczbowych. Pierwsza kolumna danych (od) określa początek przedziału, druga kolumna (do) określa jego koniec.

Tabela 3. Tabele obiektów symbolicznych zawierających dane o biurach podróży w Polsce w latach 2000-2002

Biuro podróży	Liczba klientów na imprezach turystycznych w latach 2000-2002 (w tys. osób) (A)		Liczba klientów na imprezach w kraju w latach 2000-2002 (w tys. osób) (B)		Przychody roczne w latach 2000-2002 (w mln zł) (C)		Przychody ze sprzedaży imprez krajowych roczne w latach 2000-2002 (w mln zł) (D)	
	od	do	od	do	od	do	od	do
Orbis PBP	249,51	297,28	204,881	226,50	176,89	207,00	143,9	150,10
My Travel Poland	110	126,53	40	43,88	258,05	313,07	48,548	52,473
Nowa Itaka	70	79,93	50,06	58,17	153	178,49	21,492	24,743
VING	40	49,72	23,7	25,17	105	119,80	63,56	65,25
Neckermann Polska BP	101,35	112,29	10,11	11,12	191,4	238,33	75,27	91,96
TUI Polska	95,978	98,95	2,432	2,91	188,8	202,88	74,069	86,292
Triada BP	91,43	113,64	4,33	4,80	80,585	88,97	35,14	43,66
Gromada OST	81,336	99,78	56,634	57,21	67,298	81,25	19,906	24,589
Scan Holiday Travel	80	84,14	1,9	1,98	73,4	77,79	27,95	29,02
Almatur	72	74,84	15	15,97	75	86,20	25,4	25,61
Polska BPiT								
Quo Vadis	70	75,55	68	69,20	43	43,64	50,58	53,80
Interferie	58,516	66,98	58,516	67,70	56	56,03	21,52	26,73

Źródło: opracowanie własne na podstawie rankingu biur podróży w „Rzeczpospolita” 2002; 2003).

PC2 (36.51)



Rys. 3. Konfiguracja biur podróży na płaszczyźnie wyznaczonej przez pierwszą i drugą składową główną

Źródło: opracowanie własne z użyciem pomocy programu SODAS v 1.20.

Traktując dane z tab. 3 jako tabelę danych symbolicznych, zastosowano dla nich metodę głównych składowych. Do tego celu wykorzystano pakiet *Symbolic Official Data Analysis System* (SODAS) w wersji 1.20. Rysunek 3 przedstawia konfigurację danych z tab. 3 na płaszczyźnie wyznaczonej przez pierwszą i drugą składową główną.

Należy zaznaczyć, że, w przeciwieństwie do danych „tradycyjnych”, na płaszczyźnie nie są przedstawione punkty, ale prostokąty. Wydaje się to naturalne, ponieważ reprezentują one przedziały, a nie pojedyncze wartości liczbowe.

Ładunki czynnikowe poszczególnych zmiennych przedstawia tab. 4. Wyraźnie wyodrębniono dwie składowe główne PC1 (identyfikowalna ze zmiennymi *liczba klientów na imprezach turystycznych (A)* i *liczba klientów na imprezach krajowych (B)*) oraz PC2 (identyfikowalna ze zmiennymi *przychody roczne (C)* i *przychody roczne z imprez krajowych (D)*).

Tabela 4. Ładunki czynnikowe poszczególnych zmiennych

	PC1	PC2	PC3	PC4
A	0,892	-0,357	-0,039	-0,275
B	0,718	-0,629	0,183	0,235
C	0,128	0,555	-0,384	0,118
D	0,061	0,793	0,397	-0,021

Źródło: opracowanie własne z użyciem programu SODAS v 1.20.

Struktura klas na płaszczyźnie wyznaczonej przez te dwie składowe jest dość czytelna, z trzema jednoelementowymi klasami *Orbis*, *TUI Polska* i *My Travel Poland* oraz czwartą klasą, zawierającą pozostałe biura podróży.

6. Podsumowanie

W artykule przedstawiono rozszerzenie metody głównych składowych na obiekty symboliczne, zawierające zmienne w postaci przedziałów liczbowych. Zaprezentowany przykład może świadczyć o przydatności zastosowania i czytelności omawianej metody. Należy jednak podkreślić, że przedstawione algorytmy nie umożliwiają przetwarzania danych z dowolnych obiektów symbolicznych, a jedynie z obiektów zawierających zmienne symboliczne w postaci przedziałów liczbowych, więc nie mogą być traktowane jako algorytmy uniwersalne. Co więcej, wydaje się, że zastosowanie metody głównych składowych, wymagającej obliczania wartości własnych macierzy danych, może być trudne dla danych jakościowych. Pełne zaimplementowanie metody głównych składowych i innych metod analizy czynnikowej dla obiektów symbolicznych jest zatem zagadnieniem otwartym. Szybki rozwój symbolicznej analizy danych pozwala mieć nadzieję na znalezienie sposobu na powstawanie nowych algorytmów dla danych innych niż numeryczne.

Literatura

- Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, red. H.H. Bock, E. Diday, Springer Verlag 2000.
- Diday E., *An Introduction to Symbolic Data Analysis and the SODAS Software*, J.S.D.A., International E-Journal 2002.
- Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998.
- Hair J.E., Anderson R.E., Tatham R.L., Black W.C., *Multivariate Data Analysis*, Prentice Hall, New Jersey 1998.
- Jajuga K., *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993.
- Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, red. E. Gatnar, M. Walesiak, AE, Wrocław 2004.
- „Rzeczpospolita” 2002, nr 97.
- „Rzeczpospolita” 2003, nr 108.

FACTORIAL ANALYSIS OF SYMBOLIC OBJECTS

Summary

This article presents some extensions of factorial analysis and principal component analysis onto data represented in form of symbolic object table (matrix with symbolic objects as rows). Each cell in symbolic object table can contain single quantitative value, categorical value, interval, multivalued variable, multivalued variable with weights. Vertices and Centers algorithms implementing principal component analysis on symbolic objects are described and an example of using principal component analysis on data in form of numeric intervals is presented.