PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS **nr 468 • 2017**

Taksonomia 28 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Mirosława Sztemberg-Lewandowska
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: miroslawa.sztemberg-lewandowska@ue.wroc.pl

ANALIZA NIEZALEŻNYCH GŁÓWNYCH SKŁADOWYCH

INDEPENDENT COMPONENT ANALYSIS

DOI: 10.15611/pn.2017.468.23
JEL Classification: C38

Streszczenie: Analiza głównych składowych jest metodą transformacji zmiennych pierwotnych w zbiór nowych wzajemnie nieskorelowanych zmiennych zwanych głównymi składowymi. Składowe nieskorelowane nie gwarantują niezależności ukrytych zmiennych. Składowe niezależne wyznacza się za pomocą niezależnej analizy głównych składowych (*independent component analysis*). W artykule przedstawione zostaną podstawowe podobieństwa i różnice klasycznej i niezależnej analizy głównych składowych. Walor oryginalności pracy polega na przedstawieniu przykładu zastosowania niezależnej analizy głównych składowych.

Słowa kluczowe: analiza głównych składowych, zmienne niezależne, analiza niezależnych głównych składowych.

Summary: Principal component analysis is a method of transformation of original variables in the new set of uncorrelated variables called principal components. Uncorrelated components do not guarantee the independence of variables. Independent component is determined by an independent component analysis. The paper presents the basic similarities and differences between classical and independent component analysis. The originality of the work is to present the use of an independent component analysis.

Keywords: principal component analysis, independent variables, independent component analysis.

1. Wstęp

Analiza głównych składowych została opracowana przez H. Hotellinga w 1933 r. [Harman 1975; Zakrzewska 1994]. Jest to metoda transformacji zmiennych pierwotnych w zbiór nowych nieskorelowanych zmiennych zwanych głównymi składowymi.

Analizę głównych składowych stosuje się [Górniak 2000] w przypadkach, gdy:

- celem analizy jest rozpoznanie struktury zbioru danych lub przedstawienie graficzne tego zbioru w przestrzeni dwu- bądź trójwymiarowej przy możliwie najlepszym zachowaniu relacji pomiędzy danymi, lub określenie minimalnej liczby wymiarów, za pomocą których można wyjaśnić założoną część wariancji zmiennych,
- wariancja specyficzna i wariancja wynikająca z błędu jest mała lub gdy analizuje się dużo skorelowanych zmiennych bądź korelacja między zmiennymi jest dość wysoka,
- celem rozważań jest wyznaczenie nieskorelowanych głównych składowych i zastosowanie ich w dalszych analizach wielowymiarowych (np. regresji lub dyskryminacji).

W klasycznej analizie głównych składowych wyodrębnia się nieskorelowane składowe, które nie zawsze są niezależne. Jeśli dane posiadają rozkład Gaussa, to główne składowe wyodrębnione za pomocą analizy głównych składowych są niezależne. Natomiast dane nieposiadające rozkładu Gaussa dają składowe, które są tylko nieskorelowane. Analiza niezależnych głównych składowych polega na wyznaczeniu statystycznie niezależnych składowych nawet w przypadku danych nieposiadających rozkładu normalnego [Hagai Attias, Keck 1999].

W artykule przedstawiono podstawowe podobieństwa i różnice klasycznej i niezależnej analizy głównych składowych, które zobrazowano na przykładzie empirycznym dotyczącym szkolnictwa na poziomie ponadgimnazjalnym.

2. Klasyczna i niezależna analiza głównych składowych

Model **analizy głównych składowych** można zapisać wzorem [Crawford, Lomas 1980]:

$$z_{ip} = b_{p1}s_{1i} + b_{p2}s_{2i} + \ldots + b_{pm}s_{mi} = \sum_{j=1}^{m} b_{pj}s_{ji}, \tag{1}$$

gdzie: z_{ip} – wartość p-tej zmiennej dla i-tej obserwacji $p \in \{1, 2, ..., m\}$, $i \in \{1, 2, ..., n\}$; s_{ji} – wartość j-tej głównej składowej dla i-tej obserwacji $j \in \{1, 2, ..., m\}$; b_{pj} – współczynniki głównych składowych.

W zapisie macierzowym model analizy głównych składowych przyjmuje postać:

$$\boldsymbol{Z} = \boldsymbol{B} \circ \boldsymbol{S}, \tag{2}$$

gdzie: $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m]^T$ – macierz standaryzowanych zmiennych, $\mathbf{Z}_p = (z_{1p}, z_{2p}, ..., z_{np})$; n – liczba obserwacji; $\mathbf{B} = \left[b_{pj} \right]_{m \times m}$ – macierz współczynników

głównych składowych; $\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m]^T$ – macierz głównych składowych; $\mathbf{S}_j = (s_{j1}, s_{j2}, \ldots, s_{jn})$; $p \in \{1, 2, ..., m\}$ – numer zmiennej; $j \in \{1, 2, ..., m\}$ – numer głównej składowej *T*, – znak transpozycji.

W celu wyznaczenia współczynników głównych składowych stosowany jest algorytm głównych składowych Hotellinga [Pluta 1986]. Współczynniki składowych w tej metodzie określa się w sposób iteracyjny. W pierwszym kroku ustala się współczynniki pierwszej składowej $\mathbf{S}_1$ poprzez maksymalizację udziału tej składowej w wariancji wszystkich zmiennych (W_1), tzn. maksymalizując funkcję:

$$W_1 = \sum_{p=1}^{m} b_{p1}^2 \tag{3}$$

za pomocą mnożników Lagrange'a, przy ograniczeniu $\tilde{\mathbf{R}} = \mathbf{B}\mathbf{B}^T$ (gdzie $\tilde{\mathbf{R}}$ jest macierzą kowariancji). W drugim kroku oblicza się macierz pozostałości kowariancyjnej:

$$\tilde{\mathbf{R}}_1 = \tilde{\mathbf{R}} - \mathbf{B}_1\mathbf{B}_1^T, \tag{4}$$

gdzie w miejsce $\mathbf{B}_1 = \left[b_{p1} \right]$, $p \in \{1, 2, ..., m\}$ podstawia się wartości ładunków pierwszej składowej. Określoną w ten sposób $\tilde{\mathbf{R}}_1$ podstawia się w miejsce $\tilde{\mathbf{R}}$ do równania $\tilde{\mathbf{R}} = \mathbf{B}\mathbf{B}^T$ i wyznacza się ładunki drugiej składowej głównej $\mathbf{S}_2$. W analogiczny sposób wyznacza się ładunki trzeciej i następnych składowych głównych, aż do osiągnięcia wymaganego łącznego stopnia wyjaśniania przez nie wariancji zmiennych (np. 85%) lub do momentu, gdy udział kolejnej składowej jest nie mniejszy niż wcześniej ustalona wartość (np. 5%).

Współczynnik korelacji, wykorzystywany w klasycznej analizie głównych składowych, sprawdza się jako miara niezależności składowych tylko w przypadku wielowymiarowego rozkładu normalnego. Z tego powodu w **metodzie składowych niezależnych** do badania niezależności składowych wykorzystuje się miary oparte na entropii. W statystyce entropię interpretuje się jako średnią wartość funkcji określonej na zbiorze prawdopodobieństw wszystkich możliwych realizacji pewnego doświadczenia [Liu i in. 2016]. Funkcja ta określa ilość informacji, jaką niesie pojedyncze zdarzenie. Oznaczmy przez $H(\mathbf{X})$ entropię zmiennej **X**, wówczas

$$H(\mathbf{X}) = H(\mathbf{X}) = -\sum_{i=1}^{k} p_i \log_2 p_i, \tag{5}$$

gdzie p_i to prawdopodobieństwo wystąpienia zdarzenia x_i. Entropia jest zawsze nieujemna i równa zeru tylko w takim przypadku, gdy jedno zdarzenie występuje z prawdopodobieństwem równym jedności, a pozostałe mają prawdopodobieństwa równe zeru. Natomiast osiąga wartość maksymalną w przypadku, gdy prawdopodobieństwa wszystkich zdarzeń są równe.

Aby określić stopień zależności między zmiennymi, konstruuje się miarę określaną jako wzajemna informacja $I(\mathbf{X})$, której podstawą jest entropia poszczególnych zmiennych. Wzajemna informacja obliczana jest jako suma różnic między entropiami gęstości rozkładów brzegowych zmiennych.

$$I(\mathbf{X}) = \sum_j (H(\mathbf{X}_j) - H(\mathbf{X})). \tag{6}$$

Miara ta jest modyfikacją funkcji odległości Kullbacka-Leiblera dla dwóch rozkładów prawdopodobieństwa, w której nie jest wykorzystywana entropia.

W praktyce do mierzenia zależności zmiennych stosuje się negentropię $J(\mathbf{X}_p)$, czyli miarę, która określa, jak bardzo różni się rozproszenie i koncentracja cechy o dowolnym rozkładzie od cechy o takiej samej wariancji, ale podlegającej rozkładowi normalnemu. Podstawą porównania jest rozkład normalny, ponieważ zmienna podlegająca temu rozkładowi charakteryzuje się największą entropią.

$$J(\mathbf{X}_p) = H(\mathbf{Y}_p) - H(\mathbf{X}_p), \tag{7}$$

$\mathbf{Y}_p$ jest losową zmienną podlegającą rozkładowi normalnemu o takiej samej wariancji jak $\mathbf{X}_p$. Negentropia jest nieujemna i mierzy odległość rozkładu składowej $\mathbf{X}_p$ od rozkładu normalnego.

Niezależna analiza głównych składowych, a właściwie analiza niezależnych głównych składowych (*Independent Component Analysis* – ICA) zakłada, że zmienne obserwowalne są liniową kombinacją wzajemnie niezależnych ukrytych składowych [Côme i in. 2010]:

$$\boldsymbol{X} = \boldsymbol{S} \circ \boldsymbol{A} + \boldsymbol{E}, \tag{8}$$

gdzie: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m]$ – macierz standaryzowanych zmiennych,
$\mathbf{X}_p = (x_{1p}, x_{2p}, \ldots, x_{np})$, n – liczba obserwacji,

$\mathbf{A} = [a_{pj}]_{m \times m}$ – macierz współczynników głównych składowych,

$\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m]$ – macierz głównych składowych (inaczej sygnały źródłowe), $\mathbf{S}_j = (s_{j1}, s_{j2}, \ldots, s_{jn})$,

$\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_m]$ – macierz zawierająca szumy,

$p \in \{1, 2, \ldots, m\}$ – numer zmiennej,

$j \in \{1, 2, \ldots, m\}$ – numer głównej składowej.

Ponieważ zmienne są standaryzowane, to oczywiste jest, że kolumny macierzy **X** mają średnią równą zero.

Zależność między rozkładami obserwowalnych i ukrytych zmiennych jest wyrażona równaniem [Hagai Attias, Keck 1999]:

$$f^X(\mathrm{x}) = \frac{1}{|\det(\boldsymbol{A})|} f^S(\boldsymbol{A}^{-1}x). \tag{9}$$

Równanie (9) narzuca założenie – macierz $\boldsymbol{A}$ jest nieosobliwa.

Celem metody jest wyznaczenie macierzy **W** takiej, że kolumny macierzy $\boldsymbol{S} = \boldsymbol{X} \circ \boldsymbol{W}^T$ są niezależne (czyli główne składowe są niezależne).

Niech $\boldsymbol{A}^T = \boldsymbol{P} \circ \boldsymbol{R}$, gdzie $\boldsymbol{R}$ jest macierzą rotacji ortogonalnej, oraz $\boldsymbol{Q} = \boldsymbol{P}^{-1}$, to $\boldsymbol{Y} = \boldsymbol{X} \circ \boldsymbol{Q}^T$ można zapisać $\boldsymbol{W} = \boldsymbol{R} \circ \boldsymbol{Q}$. Celem jest wyznaczenie ortogonalnej rotacji $\boldsymbol{R}$, takiej, że estymowane kolumny macierzy $\boldsymbol{S} = \boldsymbol{Y} \circ \boldsymbol{R}$ są niezależne.

Macierz odwrotna do macierzy ładunków składowych estymowana jest za pomocą np. formuły *log-likelihood*:

$$L(\boldsymbol{W}, \boldsymbol{X}) = \sum_i \sum_j \log(f^{S_j}((\boldsymbol{W}x_i)_j)) + n\log(\det(\boldsymbol{W})). \tag{10}$$

3. Przykład empiryczny

W celu zobrazowania omawianych metod czynnikowych zostanie przytoczony przykład empiryczny. Celem badania jest porównanie województw ze względu na sytuację szkolnictwa poziomu 3 ISCED. Dane pochodzą z Banku Danych Regionalnych, dotyczą 2015 roku. Zmienne uwzględnione w badaniu:

X1 – uczniowie obowiązkowo uczący się języka obcego/uczniowie liceów ogólnokształcących.

X2 – uczniowie dodatkowo uczący się języka obcego/uczniowie liceów ogólnokształcących.

X3 – absolwenci liceów ogólnokształcących/uczniowie liceów ogólnokształcących.

X4 – współczynnik skolaryzacji brutto w % (16-18 lat).

X5 – zdawalność egzaminów maturalnych w %.

X6 – uczniowie liceów ogólnokształcących/nauczyciele pełnozatrudnieni i niepełnozatrudnieni w przeliczeniu na etat.

X7 – uczniowie przypadający na 1 oddział w szkołach ogólnokształcących.

X8 – uczniowie liceów ogólnokształcących/szkoły ogólnokształcące razem.

X9 – oddziały w szkołach ogólnokształcących/szkoły ogólnokształcące razem.

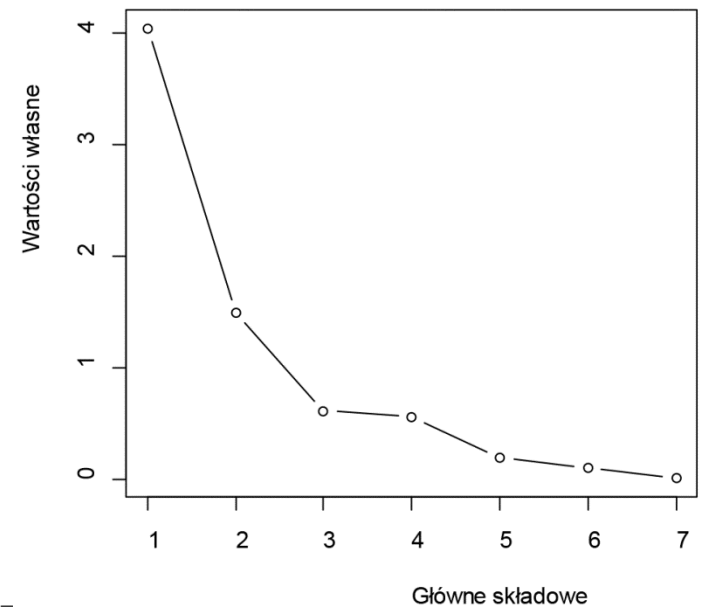
Przed przystąpieniem do procedury analizy głównych składowych dokonano selekcji zmiennych, obliczając miary adekwatności doboru każdej indywidualnej zmiennej – MSA (tabela 1).

Na podstawie MSA wyeliminowano z badania zmienne X3 i X5. Wskaźnik Kaisera-Meyera-Olkina dla pozostałych zmiennych wynosi 0,6. Na podstawie wykresu osypiska (rys. 1) za pomocą **analizy głównych składowych** wyodrębniono dwie składowe, które wyjaśniają 85% zasobu zmienności wspólnej wszystkich zmiennych.

Tabela 1. Miary adekwatności doboru każdej indywidualnej zmiennej

X1	X2	X3	X4	X5	X6	X7	X8	X9
0,441	0,567	0,337	0,423	0,399	0,533	0,463	0,515	0,520

Źródło: obliczenia własne z wykorzystaniem programu R.

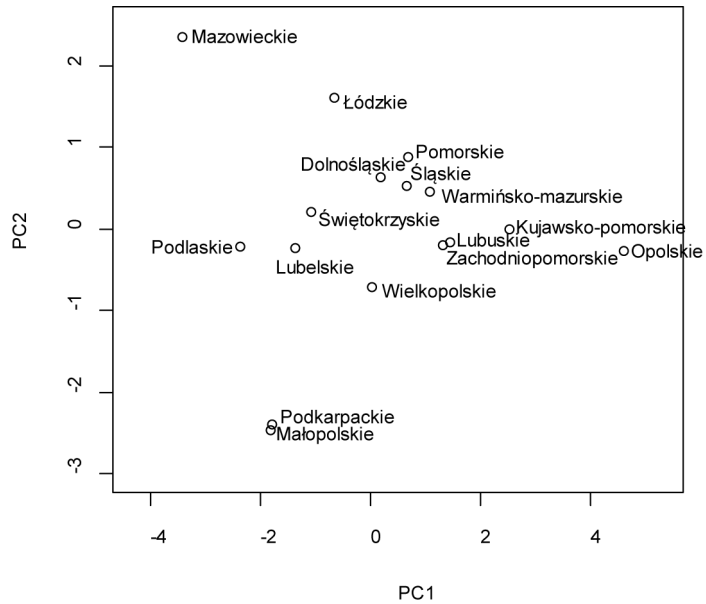


Rys. 1. Wykres osypiska

Źródło: opracowanie własne z wykorzystaniem programu R.

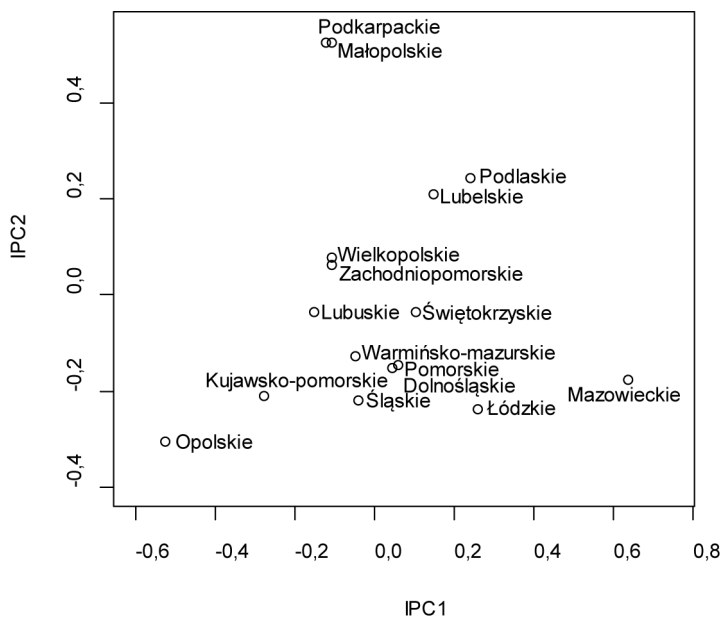
Pierwszą składową główną opisują zmienne: X7 (uczniowie przypadający na 1 oddział w szkołach ogólnokształcących), X8 (uczniowie liceów ogólnokształcących/szkoły ogólnokształcące razem), X9 (oddziały w szkołach ogólnokształcących/szkoły ogólnokształcące razem) – jest związana z zapleczem dydaktycznym szkoły. Drugą składową główną opisują zmienne: X1 (uczniowie obowiązkowo uczący się języka obcego/uczniowie liceów ogólnokształcących), X2 (uczniowie dodatkowo uczący się języka obcego/uczniowie liceów ogólnokształcących), X4 (współczynnik skolaryzacji), X6 (uczniowie liceów ogólnokształcących/nauczyciele pełnozatrudnieni i niepełnozatrudnieni). Druga składowa opisuje liczebność uczniów.

Na podstawie wyników PCA (rys. 2) można wyróżnić województwa odstające: podkarpackie i małopolskie; mazowieckie oraz łódzkie. Pozostałe województwa tworzą grupę obiektów podobnych do siebie pod względem badanych zmiennych.



Rys. 2. Wykres obiektów w przestrzeni składowych głównych PCA

Źródło: opracowanie własne z wykorzystaniem programu R.



Rys. 3. Wykres obiektów w przestrzeni składowych głównych ICA

Źródło: opracowanie własne z wykorzystaniem programu R.

Następnie przeprowadzono analizę niezależnych głównych składowych. Tutaj także wyodrębniono dwie składowe. Pierwsza składowa główna opisuje liczebność uczniów (X1, X4). Druga składowa główna jest związana z zapleczem dydaktycznym (X7, X8).

Województwami odstającymi, podobnie jak w przypadku PCA, są podkarpackie i małopolskie; mazowieckie oraz łódzkie (rys. 3). Dodatkowo można zauważyć, że podlaskie i lubelskie reprezentują podobną sytuację, podobnie jak zachodniopomorskie i wielkopolskie oraz dolnośląskie i pomorskie.

4. Zakończenie

Podstawy teoretyczne wskazują przewagę stosowania analizy niezależnych głównych składowych nad klasyczną analizą głównych składowych:

- bardziej sensowne składowe otrzymane z optymalizacji warunku niezależności (ICA) niż z maksymalizacji wariancji (PCA),
- warunek niezależności obejmuje nie tylko niezależność liniową – nieskorelowanie, ale także np. wykładniczą, wielomianową,
- ICA może potencjalnie wyodrębnić dodatkowe informacje ze zbioru danych.

Wyniki otrzymane za pomocą niezależnej analizy głównych składowych dały bardziej szczegółowy obraz zależności między zmiennymi, a także relacji między badanymi obiektami. Jednak analiza niezależnych głównych składowych daje mniejsze ładunki głównych składowych, co często prowadzi do trudności w ich interpretacji.

Literatura

Côme E., Oukhellou L, Denœux T., Aknin P., 2010, *Fault diagnosis of a railway device using semi-supervised independent factor analysis with mixing con-straints*, Pattern Analysis and Applications – PAA , s. 1-14.

Crawford I.M., Lomas R.A., 1980, *Factor Analysis – a Tool for Data Reduction*, European Journal of Marketing, vol. 14, no. 7, s. 414-421.

Górniak J., 2000, *My i nasze pieniądze*, Wydawnictwo Aureus, Kraków.

Hagai Attias, Keck W.M., 1999, *Independent factor analysis*, Neural Computation – NECO, vol. 11, no. 4, s. 803-851.

Harman H., 1975, *Modern Factor Analysis*, The University of Chicago Press.

Kim J.O., Mueller C.W., 1978, *Factor Analysis. Statistical Methods and Practical Issues*, Sage, Beverly Hills.

Liu Y., Smirnov K., Lucio M., Gougeon R.D., Alexandre H., Schmitt-Kopplin P., 2016, *MetICA: independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics*, BMC Bioinformatics, DOI 10.1186/s12859-016-0970-4.

Pluta W., 1986, *Wielowymiarowa analiza porównawcza w modelach ekonometrycznych*, PWN, Warszawa.

Zakrzewska M., 1994, *Analiza czynnikowa w budowaniu i sprawdzaniu modeli psychologicznych*, UAM, Poznań.