ROZDZIAŁ 4

ANALIZA SKŁADOWYCH GŁÓWNYCH

Po przeczytaniu tego rozdziału poznamy:

- podstawowe idee leżące u podstaw analizy składowych głównych,
- sposób określania składowych głównych,
- interpretację graficzną metody składowych głównych,
- założenia analizy składowych głównych,
- zastosowanie metody składowych głównych.

I. Wprowadzenie

Zjawiska biologiczne i medyczne cechuje wielka złożoność. Dla lekarzy każdy pacjent, każda żywa komórka stanowi jedyną i niepowtarzalną indywidualność posiadającą mnóstwo różnych własnych cech. Jeśli więc chcemy nawet w przybliżeniu opisać zjawiska biologiczne, to mamy do dyspozycji ogromną liczbę zmiennych. Tak więc w badaniach medycznych i eksperymentach biologicznych występuje bardzo duża liczba zmiennych, które powinniśmy uwzględnić w opisie naszego eksperymentu. Drugim przykładem, w którym występuje duża liczba zmiennych, są różne ankiety lub kwestionariusze. Wprawdzie im prostsza ankieta, tym chętniej i szybciej odpowiadają ankietowani. Czasami jednak na wstępnym etapie budowy ankiety możemy umieścić w niej więcej pytań powiązanych ze sobą, zwiększając niepotrzebnie rozmiar ankiety. Bywają również badania, w których analizuje się kilkaset, a nawet kilka tysięcy zmiennych. Wynikają z tego dwa przeciwstawne kryteria wyboru zmiennych:

- Z jednej strony, dla celów predykcji, chcemy jak najpełniej opisać dane powiązania. Staramy się wówczas analizować jak największą liczbę zmiennych, aby opis sytuacji był jak najbardziej wiarygodny.
- Z drugiej strony ograniczają nas koszty związane z uzyskaniem informacji od dużej liczbie zmiennych. Również przy dużej liczbie zmiennych procedury statystyczne (nawet w dobie komputerów) stają się skomplikowane, a czasami niemożliwe do wykonania. Przy dużej liczbie zmiennych wzrasta też skala trudności interpretacji. Co więcej część z tych zmiennych może okazać się nieistotna albo silnie skorelowana z pozostałymi. W „gąszczu" cech istnieje też możliwość zagubienia tych najważniejszych, a które mogą decydować o stanie zdrowia pacjenta. Powinniśmy wiec ograniczyć się w doborze zmiennych niezależnych. Zbiór mniejszy uważa się zwykle za bardziej podstawowy. Prostsza staje się też interpretacja otrzymanych zależności.

W praktyce grupowanie zmiennych jest bardzo pożądane. Spotykamy się z tym przy doborze najlepszych cech diagnostycznych. Z problemem redukcji zmiennych spotkaliśmy

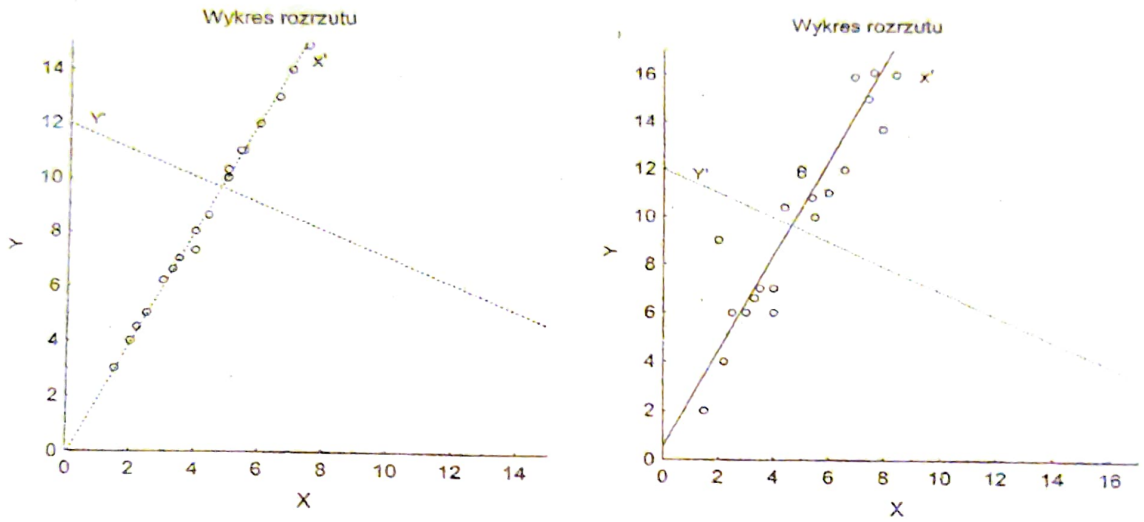
się również w poprzednim rozdziale. Wykorzystaliśmy analizę skupień do grupowania zmiennych, które były mocno skorelowane. Podobne zmienne łączyliśmy w oddzielne skupienia. Z subiektywną redukcją zmiennych spotykamy się niemal codziennie. Mówiąc, że student medycyny poradził sobie ze wszystkimi zajęciami na uczelni dzięki inteligencji i pracowitości, dokonujemy takiej subiektywnej redukcji. Udział w 15 lub 20 zajęciach na uczelni z różnorodnych przedmiotów sprowadzamy do dwóch podstawowych zmiennych: inteligencji i pracowitości. W wielu sytuacjach przypisujemy osobom wiele różnych cech na podstawie jednego lub dwóch czynników, takich jak: płeć, religia czy kolor skóry.

Jak jednak dokonać optymalnej redukcji bez znaczącej utraty informacji? Czy istnieje kompromis między opisanymi powyżej skrajnościami? Niestety, nie ma idealnej procedury statystycznej służącej do wyboru najlepszego podzbioru zmiennych. O ważności tego zagadnienia świadczy znaczna grupa analiz statystycznych proponujących różne techniki redukcji wymiarowości. Należy do nich analiza czynnikowa, analiza składowych głównych, skalowanie wielowymiarowe i opisywana wcześniej analiza skupień oraz analiza dyskryminacyjna. Analizy te musimy jednak połączyć z naszą wiedzą i doświadczeniem związanym z badanym zagadnieniem. Najbardziej popularne techniki to analiza czynnikowa i analiza składowych głównych. Obie stanowią zespół metod i procedur statystycznych pozwalających na:

- redukcję liczby zmiennych,
- wykrycie struktury i ogólnych prawidłowości w związkach między zmiennymi,
- zweryfikowanie wykrytych prawidłowości i powiązań,
- opis i klasyfikację badanych obiektów w nowych (ortogonalnych) przestrzeniach zdefiniowanych przez nowe zmienne (czynniki).

W analizie czynnikowej, jak i w analizie składowych głównych formułowane są modele matematyczne w postaci układów równań liniowych. W analizie składowych głównych jest to ortogonalne przekształcenie zmiennych obserwowalnych w nowy zbiór nieskorelowanych zmiennych (składowych). Tym samym model ten nie zakłada redukcji badanych zmiennych. Całkowita wariancja obserwowalnych zmiennych jest równa sumie wariancji składowych głównych. Natomiast w analizie czynnikowej dokonujemy dekompozycji zmiennych obserwowalnych w nowy zbiór nieskorelowanych zmiennych, wśród których wyróżniamy czynniki wspólne i czynniki swoiste. Wyodrębnienie czynników wspólnych jest głównym celem analizy czynnikowej. Generują one tak zwaną wariancję wspólną. Metody te są bardzo podobne, dlatego poświęcimy im dwa kolejne rozdziały. W niniejszym rozdziale opiszemy analizę składowych głównych, a w następnym analizę czynnikową.

Zanim przejdziemy do wzorów i precyzyjnego opisu konkretnych analiz przedstawimy graficzną ideę leżącą u podstaw tych analiz. Załóżmy, że w pewnych doświadczeniach mierzyliśmy dwie cechy dla 24 obiektów. Otrzymane wyniki tworzą układ punktów na płaszczyźnie. Graficznym odzwierciedleniem tego układu jest wykres rozrzutu przedstawiony na rys. 1 po lewej stronie.

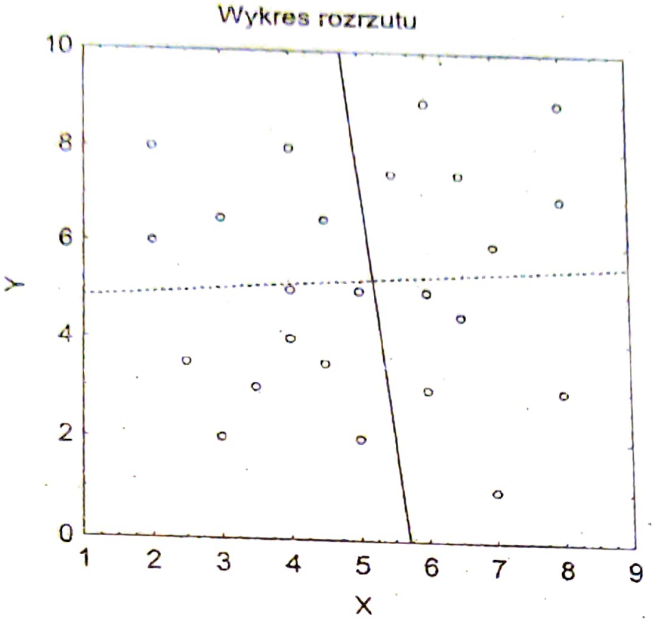


Rys. 1. Wykresy rozrzutu z zaznaczonym nowym układem współrzędnych

Przyglądając się układowi punktów nietrudno zauważyć, że punkty układają się wzdłuż przekątnej układu współrzędnych. Jeżeli zamiast dotychczasowych współrzędnych X, Y zastosuje się współrzędne w obróconym układzie X' Y' (linia przerywana), wówczas podanie współrzędnej Y' byłoby niepotrzebne. Praktycznie wszystkie punkty miałyby prawie tę samą wartość cechy Y'. Oznacza to, że do opisu zamiast dwóch zmiennych X' i Y' wystarczy, prawie bez utraty informacji, użyć jednej zmiennej X'. Mówiąc językiem statystyki, prawie cała występująca zmienność między cechami opisana jest przez zmienność cechy X', a zmienność cechy Y' możemy pominąć. Podobną sytuację przedstawia wykres po prawej stronie na rys. 1. Również i w tym doświadczeniu badane cechy są wysoce skorelowane. Możemy też użyć tylko jednej zmiennej X'. W tym przypadku redukcja zmiennych powoduje jednak pewną utratę informacji. Wynika stąd, że będąca oczywiście pod naszą kontrolą, wielkość zaniedbywanej informacji nie powinna być „zbyt duża". Jednak, jak okaże się później, jest to decyzja z natury arbitralna.

W rozważanych przykładach dokonaliśmy takiego obrotu układu współrzędnych, aby nowe osie pokrywały się z „głównymi osiami" chmury punktów na wykresie rozrzutu. Następnie dokonywaliśmy redukcji, pomijając zmienną reprezentowaną przez oś, wokół której rozrzut punktów był mały w porównaniu z rozrzutem wokół innej osi. Wynikało to z istnienia wysokiej korelacji między zmiennymi. Oczywiście, mówiąc o korelacji, będziemy zawsze myśleli o zależności liniowej. Zauważmy, że przy braku korelacji rozrzuty punktów wokół osi mogą być porównywalne i taka redukcja nie będzie miała uzasadnienia. Taka sytuacja pokazana jest na rys. 2.

Opisany powyżej przykład łączenia dwóch skorelowanych zmiennych w jeden czynnik pokazuje podstawową ideę analizy czynnikowej i składowych głównych. Możemy uogólnić podany przykład na wiele zmiennych. Obliczenia staną się bardziej złożone, ale podstawowa zasada wyrażania dwóch lub więcej zmiennych w postaci pojedynczego czynnika pozostaje taka sama. Ogólnie mówiąc, układ punktów w n wymiarowej przestrzeni może zajmować w przybliżeniu przestrzeń m wymiarową, gdzie $m < n$. Możemy zatem zbiór n zmiennych zredukować, z niewielką stratą informacji, do mniejszego zbioru m zmiennych.



Rys. 2. Wykres rozrzutu przy braku korelacji

Jak wspomnieliśmy, obecny rozdział poświęcimy analizie składowych głównych. Jest to zespół procedur statystycznych, które poprzez transformację początkowych zmiennych we wzajemne ortogonalne nowe zmienne, budują teoretyczny model opisujący strukturę zależności między badanymi cechami. Początki tej techniki pochodzą od Pearsona (1901), który stosował ją w dopasowywaniu powierzchni za pomocą metody najmniejszych kwadratów. Jednak główny rozwój tej metody zawdzięczamy pracom amerykańskiego statystyka Hotellinga (1933, 1936), który opracował ją dla analizy struktury zależności.

Oznaczmy przez p zespół początkowych zmiennych. Chcielibyśmy zredukować liczbę zmiennych, zachowując równocześnie tak dużo zmienności danych, jak to tylko jest możliwe. W analizie składowych głównych realizujemy to poprzez tworzenie nowych nieobserwowalnych zmiennych, które są kombinacją liniową zmiennych początkowych. Te nowe zmienne nazywamy składowymi głównymi. Niech Z_1 oznacza pierwszą składową główną. Wówczas zgodnie z określeniem mamy:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

gdzie $a_{11}, a_{12}, \ldots, a_{1p}$ to współczynniki wyznaczone przez początkowe zmienne.

Zauważmy, że równanie to jest podobne do równania regresji wielorakiej, gdzie Z_1 gra rolę zmiennej zależnej. Nie mylmy jednak tych równań. W analizie składowych głównych nie ma wyrazu wolnego, nie ma żadnych reszt, a co najważniejsze nie przeprowadza się rozróżnienia na zmienne zależne i niezależne.

Aby zachować tak dużo zmienności danych, jak to tylko jest możliwe, musimy wyznaczyć składową główną Z_1 o maksymalnej wariancji. Wynika stąd, że celem analizy składowych głównych jest wyznaczenie takich wartości współczynników $a_{11}, a_{12}, \ldots, a_{1p}$, aby wariancja Z_1 była tak duża, jak to tylko możliwe. W języku geometrii polega to na poszukiwaniu linii prostej, która jest najlepiej dopasowana do chmur punktów w przestrzeni. Oczywiście

musimy w tym procesie maksymalizacji zachować zdrowy rozsądek. Jeżeli bowiem współczynniki a_{1i} dążą do nieskończoności, to również wariancja Z_1 dąży do nieskończoności. Aby uciec od takiej sytuacji, maksymalizujemy wariancję Z_1 przy ograniczeniu $\sum_i a_{1i}^2 = 1$ (suma kwadratów współczynników jest równa jedności). Warunek ten jest po prostu normalizowaniem wektora współczynników. Po przekształceniach matematycznych nasz problem sprowadza się do rozwiązania układu p równań, który w zapisie macierzowym przyjmuje następującą postać:

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a}_1 = \mathbf{0}$$

gdzie: $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1p})$ wektor współczynników,

S jest macierzą kowariancji zmiennych $X_1, X_2, \ldots, X_p$,

I jest macierzą identycznościową.

Oczywiście najprostszym rozwiązaniem jest wektor zerowy $\mathbf{a}_1 = \mathbf{0}$. My jednak szukamy nietrywialnych rozwiązań. Niezerowe rozwiązanie istnieje, jeżeli macierz $(\mathbf{S} - \lambda\mathbf{I})$ nie jest macierzą odwracalną. Tak jest w sytuacji, gdy wyznacznik z tej macierzy jest równy zero, czyli $|\mathbf{S} - \lambda\mathbf{I}| = \mathbf{0}$. To ostatnie równanie nazywane jest w matematyce równaniem charakterystycznym, a λ – wartością własną macierzy **S**. Przypominamy, że podstawowe pojęcia i operacje dotyczące macierzy znajdzie Czytelnik w *Dodatku A* w tomie drugim *Przystępnego kursu statystyki*. Reasumując, aby otrzymać niezerowe rozwiązanie, musimy znaleźć wartość własną λ macierzy kowariancji **S**. Takie niezerowe rozwiązanie $\mathbf{a}_1$ nazywa się wektorem własnym odpowiadającym wartości własnej λ. Można również udowodnić, że wariancja składowej głównej Z_1 jest równa wartości własnej λ ($\mathrm{Var}(Z_1) = \lambda$). Z tych matematycznych rozważań wynika, że wszystko, co powinniśmy zrobić, to znaleźć największą wartość własną macierzy kowariancji. Współczynniki odpowiadającego jej wektora własnego to poszukiwane przez nas współczynniki $a_{11}, a_{12}, \ldots, a_{1p}$ składowej głównej. Współczynniki te wyznaczone są z dokładnością do stałej multiplikatywnej. Oczywiście w dobie komputerów wszystkie opisywane wartości wyliczają nam pakiety statystyczne.

Mimo iż pierwsza składowa jest wyodrębniona w taki sposób, że w możliwie największym stopniu wyjaśnia wariancję oryginalnych zmiennych, to jednak rzadko odtwarza ją w całości. Dlatego po znalezieniu pierwszej składowej pozostaje jeszcze trochę niewyjaśnionej zmienności. W analizie składowych głównych definiuje się dalej kolejną składową, która maksymalizuje pozostałą część zmienności. Druga składowa jest również kombinacją liniową $Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$, gdzie współczynniki $a_{21}, a_{22}, \ldots, a_{2p}$ są tak dobrane, aby wariancja Z_2 była maksymalna przy warunkach $\sum_i a_{2i}^2 = 1$ i $\sum_i a_{1i}a_{2i} = 0$.

Pierwszy z warunków jest po prostu normalizowaniem współczynników w celu ich jednoznacznego wyznaczenia. Drugi mówi o ortogonalności wektorów współczynników $\mathbf{a}_1$ i $\mathbf{a}_2$. Dzięki niemu suma wariancji kolejnych składowych głównych daje całkowitą wariancję układu. Z podobnych rozważań, jakie przeprowadzaliśmy dla pierwszej składowej, możemy pokazać, że współczynniki drugiej składowej głównej to elementy wektora własnego odpowiadającego drugiej co do wielkości wartości własnej macierzy kowariancji **S**.

W ten sam sposób możemy wyodrębnić następne składowe główne. Kolejna i-ta składowa jest kombinacją liniową $Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$, gdzie współczynniki $a_{i1}, a_{i2}, \ldots, a_{ip}$ to elementy wektora własnego, odpowiadającego i-tej co do wielkości wartości własnej macierzy kowariancji **S**. Reasumując:

- Kolejna składowa jest definiowana tak, aby maksymalizować zmienność, która nie została wyjaśniona przez poprzednią składową.
- Kolejne składowe są wzajemnie ortogonalne, tzn. są wzajemnie nieskorelowane.
- Wariancja składowej głównej Z_i jest równa i-tej co do wielkości wartości własnej macierzy kowariancji **S** ($\mathrm{Var}(Z_i) = \lambda_i$). Zatem całkowita wariancja układu jest równa $\lambda_1 + \lambda_2 + \ldots + \lambda_p$. Pozwala to zdefiniować część wariancji wyodrębnionej przez i-tą składową według wzoru:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \times 100\%$$

- Liczba wyodrębnionych w ten sposób składowych nie przekracza liczby wyjściowych zmiennych.

Jak widzimy, wartości własne macierzy kowariancji dla zmiennych początkowych odgrywają ważną rolę przy obliczaniu składowych głównych. Oprócz określania współrzędnych dają one także dobrą informację na temat wielkości wariancji wyjaśnianej przez daną składową. Informację tę możemy później wykorzystać przy decydowaniu o kolejności, według której będziemy mogli redukować wymiary oryginalnej przestrzeni zmiennych, bez straty zbyt dużego zakresu wyjściowej wariancji. Na wartościach własnych opiera się też wiele kryteriów pozwalających na wskazanie optymalnej liczby czynników w danej sytuacji.

Zgodnie z wprowadzonym określeniem każda składowa główna jest kombinacją liniową zmiennych wyjściowych zagadnienia. Interpretacja składowych głównych musi być przeprowadzana z wykorzystaniem pojęcia korelacji. Mając na uwadze ten fakt, definiowane są różne statystyki, użyteczne do celów interpretacji. Wśród nich najczęściej spotykane to współrzędne czynnikowe, nazywane także **ładunkami czynnikowymi**. Stosując język matematyczny, możemy powiedzieć, że ładunki czynnikowe są współczynnikami korelacji pomiędzy daną zmienną i składowymi. Mając zatem na uwadze ten fakt oraz cel interpretacji składowej, będziemy w naturalny sposób poszukiwać tych zmiennych, które mają najwyższe (w wartościach bezwzględnych) wartości współrzędnych czynnikowych dla danych składowych. Opisują one wkład zmiennej do poszczególnych składowych. Zauważmy także, że znak ładunków czynnikowych liczy się tylko w tym sensie, że zmienne o przeciwnych znakach ładunków dla danej składowej wnoszą odmienny wkład. Możemy jednak pomnożyć wszystkie ładunki przez -1 (zmieniając wszystkie znaki), nie wpływając w żaden sposób na wyniki.

Jak wiemy, każda składowa jest definiowana tak, aby w możliwie największym stopniu wyjaśniać zmienność oryginalnych zmiennych. Zgodnie z tym stwierdzeniem wielkości, które powinny odgrywać decydującą rolę, to miary zmienności. Istnieje wiele miar zmienności. Statystycy preferują jednak:

- Wariancję – dla pomiaru zmienności, gdy rozważamy jedną zmienną. Dla n wartości pewnej zmiennej, wyliczymy ją według wzoru: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, gdzie $\bar{x}$ oznacza ich wartość średnią.
- Kowariancję – dla pomiaru współzmienności, gdy rozważamy dwie zmienne. Dla n wartości zmiennych wyliczymy ją według wzoru: $\mathrm{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, gdzie $\bar{x}$ i $\bar{y}$ oznaczają ich wartości średnie.

Miary te zostały szeroko omówione w pierwszym tomie *Przystępnego kursu statystyki*. Dla wielu zmiennych miary te zapisujemy w postaci tzw. macierzy kowariancji. Taka macierz zawiera całą informację potrzebną do wyznaczenia składowych głównych. Przykładowa macierz kowariancji dla dwóch zmiennych pokazana jest poniżej.

$$\begin{pmatrix} s_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & s_y^2 \end{pmatrix}$$

I tu pojawia się pierwszy problem w analizie składowych głównych. Jeżeli obie zmienne wyrażają się w tych samych jednostkach i są porównywalne (tego samego rzędu), to dalsze rozważania możemy poprowadzić, opierając się na prezentowanej macierzy kowariancji. W większości jednak badań zmienne mają różne jednostki. Przykładowo wiek podajemy w latach, masę ciała w kilogramach, a składniki biochemiczne w różnorodnych jednostkach. A przecież dodawanie do siebie kilogramów, metrów i lat nie ma sensu! W takich sytuacjach musimy wykorzystać standaryzowane wersje tych zmiennych. Standaryzację przeprowadzamy według wzoru:

$$z = \frac{x_i - \bar{x}}{s_x},$$

gdzie $\bar{x}$ i s_x to średnia i odchylenie standardowe zmiennej x wyliczone z próby.

Łatwo sprawdzić, że zmienna standaryzowana ma wartość średnią równą zero i odchylenie standardowe (a zatem i wariancję) równe jedności. Macierz kowariancji dla zmiennych standaryzowanych przechodzi w tzw. macierz korelacji postaci:

$$\begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}$$, gdzie r_{xy} jest współczynnikiem korelacji

Wybór sposobu obliczania składowych jest rzeczą ważną, zwłaszcza że składowe otrzymane dla macierzy kowariancji i macierzy korelacji, ogólnie rzecz biorąc, nie muszą być takie same. Reasumując:

- Jeżeli zmienne mają różne jednostki lub są różnego rzędu, analizę składowych głównych przeprowadzamy wykorzystując macierz korelacji.
- Jeżeli analizowane zmienne są porównywalne, to wykorzystujemy macierz kowariancji. W tym przypadku wyniki analizy zależą od różnic w zakresie zmienności w obrębie

zmiennych aktywnych. Dlatego też analizę bazującą na macierzy kowariancji stosujemy tylko w przypadku, jeśli wykrycie takich różnic ma związek z rodzajem badań, które przeprowadzamy.

W takiej sytuacji wszystkie pakiety statystyczne umożliwiają wybór sposobu obliczania składowych głównych dla zmiennych: albo na podstawie macierzy korelacji, albo przy wykorzystaniu macierzy kowariancji.

Zanim przejdziemy do interpretacji geometrycznej składowych głównych oraz do omówienia pewnych praktycznych technik, utrwalimy sobie wprowadzone pojęcia na konkretnym przykładzie medycznym.

Przykład 4.1.

W losowo wybranej grupie chorych dokonano pomiaru pięciu parametrów biochemicznych:

- proerytroblasty [%],
- neutrofile [%],
- promielocyty [%],
- potas [mmol/l],
- sód [mmol/l].

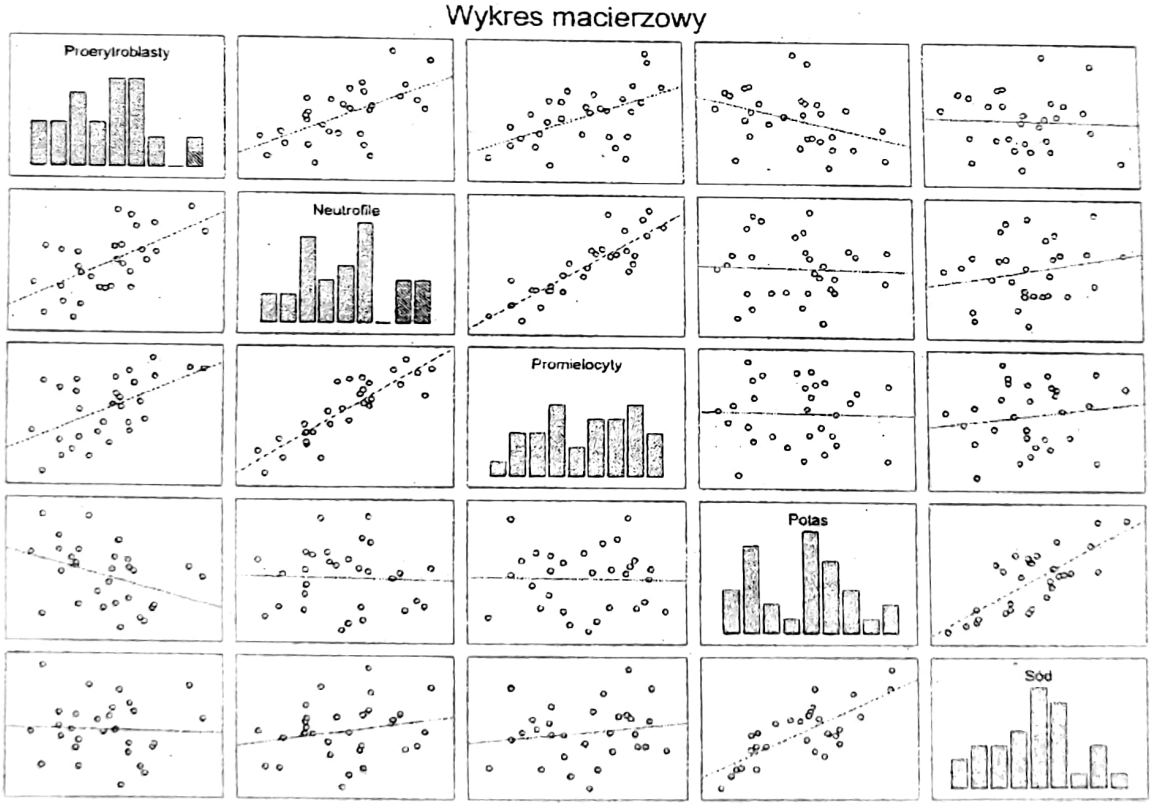
Otrzymane wyniki dla 30 osób przedstawia poniższa tabelka.

Proerytroblasty	Neutrofile	Promielocyty	Potas	Sód	Proerytroblasty	Neutrofile	Promielocyty	Potas	Sód
1,53	28,5	7,5	2,7	126,5	1,93	18,6	7,1	2,5	126,5
0,18	19,9	6,7	5,28	133	2,7	24,2	7,7	3,4	128,7
1,9	26,4	8,5	2,4	126,2	1,78	16,3	3,9	2	124
0,91	11,4	1,1	5,24	131,4	1,71	18,9	5,2	2,1	124,6
1,27	19,9	5	3,61	129,5	0,44	18,8	6,2	4,63	129,1
1,52	13	3,6	3,9	125,7	2,49	31	7,8	3,71	131,4
1,32	29,9	5,8	2,53	127,7	1,61	25,5	6,6	3,57	129,2
1,04	8,5	2,9	2,9	128,5	0,75	3,8	2,5	2,5	124,6
0,7	18,2	5,9	3,7	126,1	0,17	1,3	0	2,2	124,7
1,54	8,6	3	3,66	129,9	1,13	8,3	3,6	2,4	127,5
0,75	12,6	4,6	4,2	126,8	1,38	20,3	5,5	4,49	131
1,22	8,8	2,1	3,2	128,8	0,44	8,7	2,3	4,2	128
1,31	16,3	7	3,98	127,6	0,47	4,5	1,1	3,53	127,1
0,64	0	1,6	3,96	127,7	1,4	15,1	4,8	1,8	123
0	9,9	3	4,12	127,5	0,71	14,2	6,8	3,79	128,9

Ponieważ zmienne mają różne jednostki, więc zgodnie z wcześniejszymi rozważaniami, analizę składowych głównych przeprowadzimy na podstawie macierzy korelacji. Dla prezentowanych danych przyjmuje ona postać:

$$\mathbf{R} = \begin{array}{c} \begin{matrix} \text{Proerytro.} & \text{Neutrof.} & \text{Promiel.} & \text{Potas} & \text{Sód} \end{matrix} \\ \begin{pmatrix} 1 & \mathbf{0,61} & \mathbf{0,55} & \mathbf{-0,37} & -0,03 \\ \mathbf{0,61} & 1 & \mathbf{0,87} & -0,03 & 0,23 \\ \mathbf{0,55} & \mathbf{0,87} & 1 & -0,02 & 0,17 \\ \mathbf{-0,37} & -0,03 & -0,02 & 1 & \mathbf{0,76} \\ -0,03 & 0,23 & 0,17 & \mathbf{0,76} & 1 \end{pmatrix} \end{array} \begin{matrix} \text{Proerytro.} \\ \text{Neutrof.} \\ \text{Promiel.} \\ \text{Potas} \\ \text{Sód} \end{matrix}$$

Istotne korelacje zaznaczone są czcionką pogrubioną. Zauważamy wysoką korelację między trzema pierwszymi zmiennymi. Również dwie ostatnie zmienne są istotnie skorelowane na poziomie 0,76. Pozostałe korelacje są względnie niskie. Zatem wydaje się, że w macierzy tej występuje jakaś wyraźna struktura, która będzie miała wpływ na postać składowych głównych. Graficznie sytuację tę pokazuje rys. 3.



Rys. 3. Macierzowy wykres rozrzutu dla danych z przykładu 4.1

Wartości własne wyznaczone dla analizowanych danych przedstawione są w poniższej tabelce. Jak wiemy, i-ta wartość własna jest równa wariancji i-tej składowej głównej Z_i. Pozwoliło to zdefiniować część całkowitej wariancji wyodrębnionej przez i-tą składową. Ta ostatnia wartość podana jest w drugiej kolumnie. Kolumna trzecia i czwarta pokazują skumulowane wartości własne i skumulowany procent wariancji.

Wartość własna λ_i	Procent wariancji	Skumulowana wartość własna	Skumul. % wariancji
2,414	48,279	2,414	48,279
1,845	36,891	4,259	85,170
0,458	9,155	4,716	94,325
0,170	3,405	4,887	97,730
0,113	2,269	5,000	100,0000

Dla naszych danych składowa odpowiadająca największej wartości własnej (2,414) wyjaśnia około 48,3% całkowitej wariancji. Druga składowa odpowiadająca drugiej wartości własnej (1,845) wyjaśnia około 36,9% całkowitej wariancji itd. Każda zmienna standaryzowana ma wariancję równą jedności, więc całkowita wariancja pięciu zmiennych wynosi 5. Ponieważ całkowity rozrzut zawarty w danych nie może się zmienić, więc skumulowana wartość własna jest również równa 5. Zatem przy analizie macierzy korelacji suma wartości własnych (zacieniowana komórka) jest równa liczbie zmiennych, na podstawie których były obliczane.

Następnie wyliczamy wektory własne odpowiadające znalezionym wartościom własnym. Są one bowiem poszukiwanymi współczynnikami składowych. Po wykonaniu obliczeń otrzymujemy:

Wartość własna Składowe główne

$\lambda_1 = 2{,}414$ $Z_1 = 0{,}517\text{Proerytro.} + 0{,}605\text{Neutrofile} + 0{,}588\text{Promiel.} - 0{,}102\text{Potas} + 0{,}101\text{Sód}$

$\lambda_2 = 1{,}845$ $Z_2 = -0{,}204\text{Proerytro.} + 0{,}096\text{Neutrofile} + 0{,}085\text{Promiel.} + 0{,}695\text{Potas} + 0{,}677\text{Sód}$

$\lambda_3 = 0{,}458$ $Z_3 = -0{,}747\text{Proerytro.} + 0{,}258\text{Neutrofile} + 0{,}470\text{Promiel.} - 0{,}067\text{Potas} - 0{,}389\text{Sód}$

$\lambda_4 = 0{,}170$ $Z_4 = 0{,}365\text{Proerytro.} - 0{,}311\text{Neutrofile} + 0{,}207\text{Promiel.} + 0{,}656\text{Potas.} - 0{,}545\text{Sód}$

$\lambda_5 = 0{,}113$ $Z_5 = 0{,}019\text{Proerytro.} + 0{,}679\text{Neutrofile} - 0{,}619\text{Promiel.} - 0{,}267\text{Potas} - 0{,}288\text{Sód}$

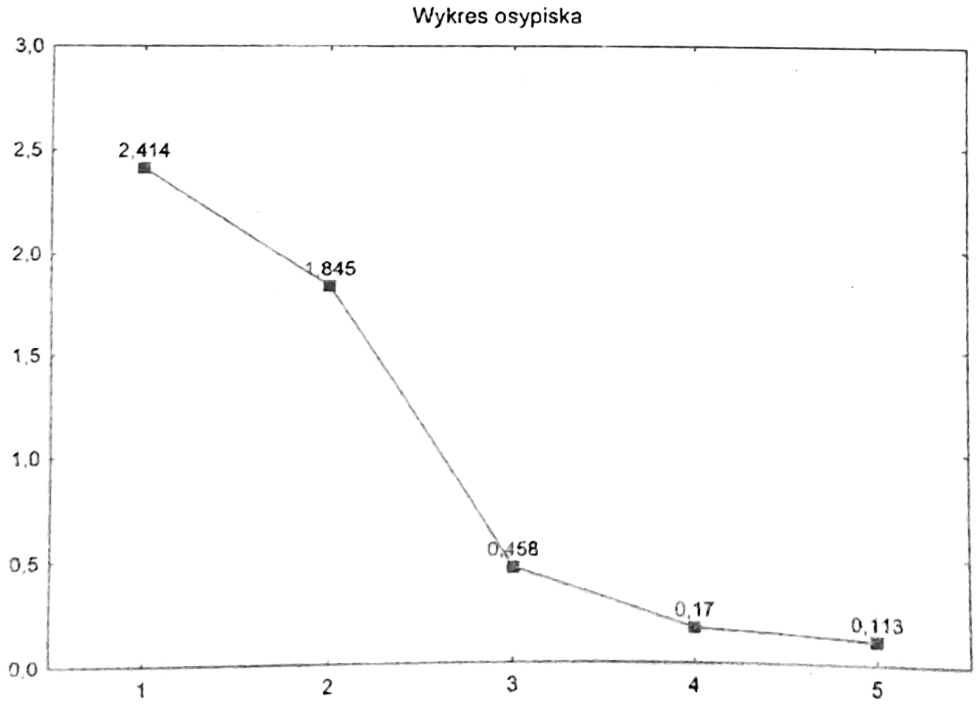
Uważny Czytelnik zapyta: Co zyskaliśmy? Startowaliśmy z pięcioma zmiennymi i po żmudnych obliczeniach otrzymaliśmy pięć (może bardziej skomplikowanych) zmiennych Z_1, Z_2, Z_3, Z_4, Z_5 (składowe). Jednak coś zyskujemy.

Po pierwsze, nowe zmienne (składowe) są ortogonalne w stosunku do siebie, tzn. są wzajemnie nieskorelowane. Takie składowe możemy wykorzystać w dalszej analizie badawczej, np. w analizie dyskryminacyjnej lub w analizie regresji wielorakiej, gdzie wymagane jest założenie o braku współliniowości.

Po drugie, każda kolejna składowa wyjaśnia coraz mniejszą część zmienności oryginalnych zmiennych. W jakimś momencie będziemy określać składową, która wyjaśnia znikomą część zmienności. Wróćmy na moment do powyższej tabeli prezentującej skumulowany procent wariancji. W naszym przykładzie okazuje się, że pierwsze dwie składowe wyjaśniają 85,2% całkowitej zmienności. Zaś trzy składowe wyjaśniają ponad 94% Pozostałe składowe wyjaśniają 6% albo mniej. Wygląda na to, że czwarta i piąta składowa są zbyteczne. Jeżeli tak, to stosując w dalszych analizach tylko „najważniejsze" składowe, dokonaliśmy redukcji zmiennych. Zamiast pięciowymiarowej przestrzeni, przy zgodzie na „niewielką" utratę informacji, wystarczy nam przestrzeń trójwymiarowa.

Czy możemy jednak pominąć pozostałe składowe? Jeżeli tak, to ile? I tu pojawia się drugi problem w analizie składowych głównych. Nie ma żadnego obiektywnego kryterium redukcji wymiarów. Wśród różnych propozycji trzy zyskały akceptację i są stosowane w praktyce.

- W pierwszym kryterium brany jest pod uwagę procent wariancji wyjaśnionej przez daną składową. Jeśli dla pierwszych dwóch, trzech składowych suma ich wariancji stanowi znaczną część wariancji wszystkich zmiennych obserwowanych, to na takiej liczbie składowych poprzestajemy. Jeśli zaś nie, to składowe wyznacza się tak długo, aż suma ich wariancji przekroczy pewną wartość, np. 75%. W literaturze (Grabiński 1992) spotykamy również ograniczenie 80% lub nawet 90%. Według tego kryterium pozostawiamy, dla danych z naszego przykładu, dwie pierwsze składowe.
- Drugie kryterium pochodzi od Kaisera (1960). Jak wiemy, standaryzowane zmienne wejściowe mają wariancję równą jeden. Zatem nowe zmienne (składowe) również powinny wnosić do opisu przynajmniej tyle samo. Stąd według kryterium Kaisera wykorzystuje się tylko te składowe, którym odpowiadają wartości własne większe od 1. W naszym przykładzie okazuje się, że tylko dwie wartości własne są większe od 1 i pozwalają na wyjaśnienie około 85% całkowitej zmienności.
- Trzecie kryterium opiera się na tzw. wykresie osypiska (Cattell 1966). Jest to prosty wykres liniowy, pokazujący kolejne wartości własne. Cattell proponuje odszukanie miejsca, od którego na prawo występuje łagodny spadek wartości własnych. Na prawo od tego punktu przypuszczalnie znajduje się tylko „osypisko czynnikowe". Jest to miejsce, w którym przyrost informacji doznaje załamania. Zatem nie powinniśmy uwzględniać więcej czynników niż te znajdujące się po lewej stronie tego punktu. Sam punkt, od którego zaczyna się osypisko, niektórzy autorzy uwzględniają przy wyznaczaniu liczby składowych, inni zaś pomijają ten punkt.



Rys. 4. Wykres osypiska dla danych z przykładu 4.1

Dla danych z naszego przykładu wykres osypiska pokazuje rys. 4 (powyżej). Na rysunku widzimy, że osypisko zaczyna się od trzeciej składowej. Jak wiemy, te pierwsze trzy składowe wyjaśniają ponad 94% całkowitej zmienności.

Które kryterium stosować? Kryterium Kaisera czasami wybiera zbyt mało składowych. Poza tym możemy go stosować, kiedy analiza jest oparta na zmiennych standaryzowanych. Dla zmiennych niestandaryzowanych, kiedy wariancja może być o wiele większa od 1, kryterium to nie powinno być stosowane. Z kolei test osypiska czasami wybiera zbyt wiele składowych. W praktyce ważnym aspektem jest to, na ile rozwiązanie poddaje się interpretacji. Dlatego zalecamy rozważenie kilku rozwiązań z większą lub mniejszą liczbą składowych, a następnie wybranie tego, które wydaje się najbardziej „sensowne". Czasami też różne względy mogą przemawiać za włączeniem dodatkowych składowych do analizy. W naszym przykładzie dołączenie nowych zmiennych jest chyba zbyteczne. Pomijając bowiem dwie ostatnie składowe, możemy powiedzieć, że zachowujemy 5,6% margines błędu.

Ponieważ opisane kryteria są bardzo subiektywne, aprobujemy więc tylko wyniki poparte przez co najmniej dwa kryteria, traktując pozostałe jako prawdopodobne hipotezy.

Interpretacja składowych

Składowe główne dla danych z naszego przykładu zapisaliśmy powyżej w postaci pięciu kombinacji liniowych. W literaturze spotykamy inny zapis – macierzowy. Współczynniki składowych zapisywane są w postaci macierzy **Z** tak, że każda i-ta kolumna zawiera współczynniki i-tej składowej głównej. Macierz współczynników składowych głównych dla danych z naszego przykładu przedstawiona jest poniżej.

Współczynniki czwartej składowej Z_4

$$Z = \begin{pmatrix} 0{,}517 & -0{,}204 & -0{,}747 & 0{,}365 & 0{,}019 \\ 0{,}605 & 0{,}096 & 0{,}258 & -0{,}311 & 0{,}679 \\ 0{,}588 & 0{,}085 & 0{,}470 & 0{,}207 & -0{,}619 \\ -0{,}102 & 0{,}695 & -0{,}067 & 0{,}656 & -0{,}267 \\ 0{,}101 & 0{,}677 & -0{,}389 & -0{,}545 & -0{,}288 \end{pmatrix}$$

Znaki i wartości współczynników a_{ij} składowych mówią nam o sposobie i wielkości wpływu i-tej zmiennej na j-tą składową. Możliwa jest także dokładniejsza i częściej stosowana interpretacja z wykorzystaniem pojęcia korelacji. Wynika ona ze spostrzeżenia, że kowariancja między i-tą zmienną a j-tą składową Z_j jest równa $\lambda_j a_{ji}$, gdzie (jak w całym rozdziale) λ_j to j-ta wartość własna a, a_{ji} współczynniki j-tej składowej. Wystarczy teraz podzielić tę wartość przez odchylenie standardowe, a otrzymamy współczynnik korelacji i-tej zmiennej z j-tą składową. W przypadku, gdy składowe główne były wyznaczane na podstawie macierzy korelacji, współczynnik korelacji między i-tą zmienną a j-tą składową jest postaci $\sqrt{\lambda_j}\, a_{j_i}$. Otrzymujemy w ten sposób tzw. ładunki czynnikowe, które możemy

zapisać w postaci macierzowej. Macierz ładunków czynnikowych dla danych z naszego przykładu przedstawiona jest poniżej.

	Z_1	Z_2	Z_3	Z_4	Z_5
Proerytro.	0,803	- 0,277	0,505	0151	0,006
Neutrof.	0,940	0,130	0,174	– 0,128	0,229
Promiel.	0,914	0,115	0,318	0,085	- 0,209
Potas	– 0,159	0,944	–,0,045	0,0271	- 0,090
Sód	0,157	0,920	- 0,263	- 0,225	- 0,097

Kolumna współczynników korelacji zmiennych z czwartą składową Z_4 równych $\sqrt{\lambda_4}a_{4i}$

Jak widzimy, pierwsza składowa Z_1 jest najbardziej skorelowana ze zmiennymi. Nie powinno nas to dziwić, składowe są bowiem wyodrębniane kolejno i wyjaśniają coraz to mniej całkowitej wariancji. Zmienne Neutrofile, Promielocyty i Proerytroblasty mają wysokie ładunki czynnikowe (0,940; 0,914 i 0,803) z pierwszą składową. Z kolei zmienne Potas i Sód mają wysokie ładunki czynnikowe z drugą składową. Pozostałe składowe wydają się nieistotne. Ten wynik koresponduje z uwagami, które sformułowaliśmy wcześniej, omawiając współczynniki korelacji. Z faktu, że składowe Z_1 i Z_2 są ortogonalne wynika, że zmienne Potas i Sód reprezentują oddzielny zestaw parametrów biochemicznych. Ze spostrzeżeń tych wynika, że pierwszą składową możemy nazwać „mielogramem", a drugą „gospodarką kationami".

Ładunki czynnikowe (współczynniki korelacji) podniesione do kwadratu możemy interpretować jako udział wyjaśnionej wariancji (współczynniki determinacji). Ładunek czynnikowy dla Neutrofili i Z_1 jest równy 0,940. Zatem $(0,940)^2 = 0,8836$, co oznacza, że 88% wariancji zmiennej Neutrofile jest wyjaśnione przez pierwszą składową „mielogram". Składowa druga dodaje $(0,130)^2$, tj. 2%, a dalsze składowe pozostałe 10%. Suma wynosi 100%, ponieważ cały rozrzut jest wyjaśniony w całości przez pięć składowych. Sumę kwadratów ładunków czynnikowych w danym wierszu nazywamy **zasobem zmienności wspólnej** (ang. *communality*). Jest to część wariancji zmiennej wyjaśniona przez składowe. Jeżeli uwzględniamy wszystkie składowe, to zasób zmienności wspólnej jest zawsze równy 100%. W naszym przykładzie, gdy uwzględnimy tylko dwie pierwsze składowe, zasób zmienności dla Neutrofili jest równy $(0,940)^2 + (0,130)^2 = 0,9005$, co oznacza, że ponad 90% wariancji tej zmiennej jest wytłumaczone przez pierwsze dwie składowe. Podobnie możemy pokazać, że 72% wariancji zmiennej Promieloblasty jest wytłumaczone przez pierwsze dwie składowe. Analogiczne obliczenia możemy przeprowadzić dla pozostałych zmiennych.

II. Interpretacja geometryczna

Jak wiemy, i-ta składowa jest kombinacją liniową $Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$, gdzie współczynniki $a_{i1}, a_{i2}, \ldots, a_{ip}$ to elementy wektora własnego odpowiadającego i-tej co do wielkości wartości własnej macierzy kowariancji **S**. Zestawiając obok siebie wszystkie składowe, otrzymamy zestaw kombinacji liniowych ([1]). Wykorzystując język

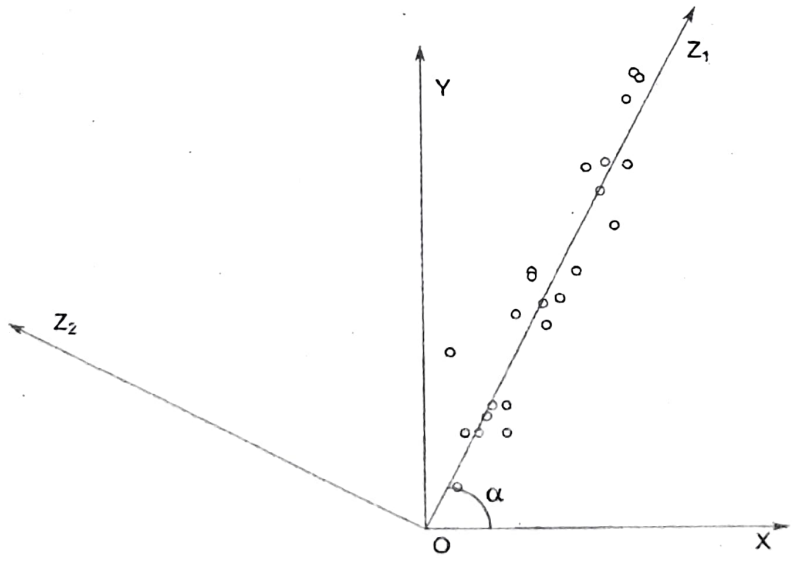
matematyki, możemy to zapisać w postaci równania macierzowego ([2]). Matematycy lubią stosować zwięzłe zapisy. I tak, oznaczając macierz współczynników pogrubioną literą **A**, a wektory składowych i zmiennych przez pogrubione **Z** i **X**, otrzymujemy prosty zapis macierzowy [3]. Przypominamy, że proste wprowadzenie do algebry macierzy znajdzie Czytelnik w pozycji *Przystępny kurs statystyki. Tom 2* (rozdział: *Dodatek A: Wektory i macierze*).

[1] [2] [3]

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Z_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \\ Z_p &= a_{p1}X_1 + a_{p2}X_p + \ldots + a_{pp}X_p \end{aligned} \Rightarrow \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} \cdot \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \Rightarrow \mathbf{Z} = \mathbf{A} \cdot \mathbf{X}$$

Po wprowadzeniu umownych oznaczeń podstawowe równanie w metodzie składowych głównych jest transformacją liniową postaci $\mathbf{Z} = \mathbf{A} \cdot \mathbf{X}$. Ponieważ kolejne składowe mają być wzajemnie ortogonalne, więc macierz **A** musi być macierzą ortogonalną. A wówczas równanie $\mathbf{Z} = \mathbf{A} \cdot \mathbf{X}$ przedstawia szczególny przypadek transformacji liniowej, a mianowicie obrót. Dla przybliżenia tego faktu przyjrzyjmy się dokładnie sytuacji dwuwymiarowej.

W przestrzeni dwuwymiarowej macierz ortogonalna **A** przyjmuje postać $\mathbf{A} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$. Oznacza to, że nowy układ OZ_1Z_2 powstaje w wyniku obrotu układu OXY o kąt α. Sytuację tę przedstawia rys. 5.



Rys. 5. Obrót układu współrzędnych

Związek pomiędzy nowymi a starymi współrzędnymi wyrażony jest wzorem:

$$Z_1 = X \cdot \cos\alpha + Y \cdot \sin\alpha$$

$$Z_2 = -X \cdot \sin\alpha + Y \cdot \cos\alpha$$

Zgodnie z wcześniejszymi rozważaniami położenie osi Z_1, czyli pierwszej składowej, jest tak dobrane, aby stanowiła „główną oś" chmury punktów w układzie OXY. Wówczas rzuty punktów na tę oś będą miały największy rozrzut. Pierwsza składowa ma bowiem wyjaśniać największą część zmienności oryginalnych zmiennych. Druga składowa musi być (jak to widać na rysunku) prostopadła do pierwszej składowej.

Uogólniając, możemy powiedzieć, że układ składowych głównych w przestrzeni n-wymiarowej powstaje w wyniku obrotu układu oryginalnego opisującego obiekty wielowymiarowe. Oczywiście początek tego układu pokrywa się z punktem centralnym reprezentowanym przez średnie wartości wszystkich zmiennych (środek ciężkości). Obrót ten następuje w taki sposób, aby kolejne osie wyjaśniały coraz mniejszy odsetek wariancji oryginalnych zmiennych.